“华为杯”第十五届中国研究生

数学建模竞赛

题 目　　对恐怖袭击事件记录数据的量化分析

摘　　要：

随着全球非常规突发事件频发，公共安全领域越来越受到重视。恐怖袭击作为一种非常规突发事件，引起了全球高度关注。近年来，全球恐怖袭击活动有增无减，就我国而言，恐怖袭击已经从南疆地区发展到了全疆，并有向全国蔓延的趋势。因此，研究恐怖袭击事件危害程度、主要原因、时空特征、蔓延趋势等特点，对全球反恐活动和政策有非常重大的意义。本文主要运用了主客观组合赋权结合灰色综合评价对恐怖袭击事件危害性分级；利用 OPTICS 聚类算法对恐袭事件进行聚类分析；运用 XGBoost 算法对每个典型事件按嫌疑程度给 5 个恐怖组织或个人排序；用 Lorenz 曲线、基尼系数与帕累托分析法研究不同变量的空间聚集性与热点；并提出 lightgbm-multilogistic 串行模型以挖掘恐怖袭击事件中死亡人数影响因素，以及基于 ITSE 方法对恐怖袭击事件数增长率建立 ARMA 干预模型，从而验证 911 事件之后反恐活动效果并研究恐袭事件数的周期性特征。

问题一要求依据危害性对恐怖袭击事件进行分级。考虑到恐怖袭击事件危害程度评价的主观性和各指标的内在联系及权重随时间渐变性的客观特点，本文基于矩估计方法将主观赋权法（层次分析 AHP、序关系分析 GI）和客观赋权法（熵权法、标准差和平均差最大化法）进行集成，以获取最优指标权重。以此为基础，本文进一步使用灰色综合评价方法对恐怖袭击事件危害性进行评分，同时使用聚类分析对得分划分等级。结果显示，2001 年的 911 事件是 1998-2017 年期间危害级别最高的恐怖袭击事件。

问题二要求依据恐怖袭击事件的特征发现其组织者。本文首先利用对聚类簇的形状没有偏倚、不需提前设定聚类个数且对参数设置不太敏感的 OPTICS 算法对恐怖袭击事件进行聚类分析，对同一类别内的事件认为是由同一恐怖组织或个人发起的，而不同类别的恐怖事件则不是，最终将所有恐怖袭击事件聚为 28 类，即这些事件可能由 28 个不同的恐怖组织或个人发起。接着从恐怖袭击发起数量

和危害程度进行考量，构建了恐怖组织或个人的危害性的综合评分方法，通过对得分的排序找出了危害性最大的 5 个恐怖组织或个人。最后利用 XGBoost 算法对每个典型事件按嫌疑程度给 5 个恐怖组织或个人排序。

问题三要求分析近三年来的数据，从多角度探究未来反恐态势。本文首先分析了全球恐怖袭击事件时空分布的总体特征，从时间维度发现近年来全球恐怖袭击事件数量和成功率均呈现持续下降趋；从空间维度发现恐袭事件的发生在空间分布上有很强的聚集性。本文接着进一步同时从时间和空间的维度研究了恐怖主义的蔓延趋势，发现近年来全球恐怖主义活动空间总体受到一定打压，但在个别区域有进一步蔓延风险。之后本文利用 Lorenz 曲线、基尼系数与帕累托分析法研究了不同变量的空间聚集性与热点，由于国家与地区等区域因素的空间聚集性十分显著，本文进一步从不同角度出发挖掘了恐袭事件在不同区域的特点。最后基于前文的研究结论，本文给出了对未来恐怖主义发展形势的预测与关于未来反恐部署的建议。

问题四要求挖掘附件 1 数据的其他价值并给出模型及方法。本文主要提出了两个方向：其一，通过问题一的探索，我们发现死亡人数成为事件危害性权重最高的指标，同时观察数据可知，近年来恐怖袭击事件表现出高死亡率的趋势，2017 年死亡人数达 26445 人。因此我们尝试从数据中找到导致高死亡的关键因素，以协助反恐政策制定，及时遏制高死亡率。针对该问题，我们提出建立有监督的机器学习 lightgbm 模型与多分类 logistic 串行融合的方法，在保证模型分类效果的同时，还能够得到各个特征重要性，挖掘主要影响因素。其二，由问题一的结论可知，911 事件是目前恐怖袭击中危害性最高的事件，自此之后，国际社会和世界各国纷纷加大反恐力度，与此同时，恐怖袭击事件也走向了新的特征和新的趋势。因此我们认为可以研究 911 事件对恐怖袭击所引起的新变化，并提出利用基于 ITSE 方法对恐怖袭击事件数增长率建立 ARMA 干预模型，从而验证 911 事件之后反恐活动效果及发现恐袭事件数的周期性特征。

关键词：恐怖袭击、主客观集成赋权、OPTICS 聚类、XGBoost 集成学习、Lorenz 曲线、帕累托分析法、lightgbm-Multi logistic、ARMA 模型

目录

一、 问题背景与重述

1.1 问题背景

近年来，全球恐怖袭击日益呈现出全球化、长期化、扩大化和复杂化的趋势，世界各国的反恐工作面临严峻考验。与此同时，随着信息技术的发展，统一完整的全球恐怖袭击信息共享数据环境逐步搭建完成。通过对这些数据进行有效分析，有助于国际社会更清晰的理解全球恐怖主义发展趋势、掌握恐怖组织的行为特征，从而对各国政府面对恐怖袭击的预防和应急管理能力的提高以及各个国家及区域间的联合反恐工作的开展具有十分重要的意义。

1.2 问题重述

题目给出了来自全球恐怖主义数据库（GTD）中 1998-2017 年世界上发生的恐怖袭击事件的记录，要求利用该数据解决以下几个问题：

题目 1：利用题目所给数据，建立一套针对恐怖袭击事件的量化分级标准，将事件按照危害程度从高到低依次划分为一至五级。之后，利用该标准对题目所给数据中的全部事件进行分级，并选出近二十年来危害程度最高的十大恐怖袭击事件。

题目 2：首先，针对 2015、2016 年尚未有组织或个人宣称负责的恐怖袭击事件，将可能是同一个恐怖组织或个人在不同时间、不同地点多次作案的若干案件归为一类，对应的未知作案组织或个人标记不同的代号。接下来，按照之前标记的组织或个人的危害性从大到小选出其中的前 5 个。最后，对给定的恐袭事件按嫌疑程度对 5 个嫌疑人进行排序。

题目 3：依据题目所给数据并结合因特网上的其他有关信息，研究近三年来恐怖袭击事件发生的主要原因、时空特性、蔓延特性、级别分布等规律，进而分析研判下一年全球或某些重点地区的反恐态势，在此基础上，提出对未来反恐斗争的见解和建议。

题目 4：利用题目所给数据，进行进一步的探讨和研究。

二、 问题分析

针对问题一需要依据危害性对恐怖袭击事件的进行分级。由于题目所给数据维度较高，且各变量间存在较高的相关性，因此本文选择利用因子分析的方法在对数据进行降维的同时通过旋转使因子具有更鲜明的实际意义。在计算权重矩阵时，层次分析、序关系分析等赋权法具有较强的主观性；而客观赋权法往往又只考虑了数据本身的特性，很可能因为数据的偏倚而导致权重不准确，因此可以选择将主观赋权法与客观赋权法结合，把定性评价转化为定量评价，使得评价标准、影响因素的模糊性得以体现，同时加强权重系数的鲁棒性。最后，在计算样本得分时，可以使用灰色综合评价法的关联分析，并利用分布或者聚类方法对其划分

等级。

针对问题二需要根据恐怖袭击事件的特征寻找嫌疑人。由于相同组织或个人发动的袭击往往具有相同的特征，因此首先可以对恐袭事件进行聚类。因为OPTICS 具有聚类簇的形状没有偏倚、不需提前设定聚类个数且对参数设置不太敏感等优点较为适合本题的情况，所以选择运用该算法得到聚类结果。接着应该从多角度全面的度量聚类得到的恐怖组织或个人的危害性。由于问题一已经得到恐袭事件的危害程度故可从恐怖袭击发起数量和危害程度两个角度出发构建综合评分模型，并基于该模型的得分进行排序找出危害性最大的 5 个恐怖组织或个人。最后在得到 5 个嫌疑人名单后需要判断它们在给定事件中的嫌疑程度，这可以看做一个多分类问题，XGBoost 算法往往可以在这类问题中取得较好的效果。

针对问题三需要从多维度研判未来全球或某些重点地区的反恐态势。从时间上可以对近年来全球恐怖袭击事件数量、成功率和等级进行探究。从空间上可以分析恐袭事件的数量和等级在空间分布上的特性。从时间加空间的角度可以寻找近年来恐怖主义的蔓延趋势。之后可尝试利用 Lorenz 曲线、基尼系数与帕累托分析法结合进一步挖掘不同变量的空间聚集性与热点。

针对问题四需要挖掘出附件 1 数据中的更多价值。首先附件中包含每起恐怖袭击事件的详细信息，从时间、区域、袭击方式、袭击目标到死伤人数、索赔金额、组织者等等；问题三已经从时空时空、蔓延趋势、原因分析等维度对其进行了分析，这里可以进一步从其他维度分析：研究恐怖袭击事件袭击方式、袭击目标的转变；研究不同组织者的恐怖袭击特点；研究危害性程度较高指标的影响因素等等。这些研究都能够有助于国家或者政府了解恐怖袭击事件态势，从而为反恐活动提供指导性建议。

三、 基本假设及符号说明

3.1 基本假设

（1）假设气候、自然灾害因素对恐怖袭击事件的发生无影响；
（2）在进行恐怖袭击事件分级时，涉及到金额的衡量，此时假设不同时间货币的购买力相同；
（3）假设同一恐怖组织或个人发起的恐怖事件具有类似的特征，同时该特征也不会因为时间的变动而变动。

3.2 符号说明

表 3.1　符号说明

符号	说明
ϵ	邻域半径
M	阈值
n_estimator	树的棵数
eta	学习效率
min_child_weight	最小叶子权重和
max_depth	树的最大深度
gamma	节点分裂所需的最小损失函数下降值
Subsample	每棵树随机采样比例
Gini	基尼系数
m	评价指标数量
n	样本数量

四、问题一的建模与求解

4.1 模型的构建

4.1.1 解题思路

对灾难性事件比如地震、交通事故、气象灾害等等进行分级有助于社会国家对其进行管理和控制。 通常的分级一般采用主观方法， 由权威组织或部门选择若干个主要指标，主观规定分级标准，如我国交通事故等级划分，主要按照人员伤亡和经济损失程度划分。

但是恐怖袭击事件的危害性不仅仅取决于这两方面，还与发生的时机、地域、对象等诸多因素有关，因此对恐怖事件划分等级的核心在于如何对各个指标进行赋权重。目前，评价指标权重的确定方法主要有主观赋权法、客观赋权法，主观赋权法主要依靠专家的经验和主观性来确定评价指标的权重及排序；客观赋权法则是根据原始数据之间的联系，通过一定的数学理论方法来确定指标的权重，反映了评价指标权重与原始数据变动的关系。本章节将使用主观赋权法和客观赋权法基于矩估计进行集成新权重，该方法具有同时兼顾事件的主观不可忽视性、待评价对象各指标的内在联系及权重随时间渐变性的特点。并以此为基础，进一步使用灰色综合评价方法对事件危害等级进行划分。

本章节使用的的主观赋权法有：改进的层次分析法（AHP）、序关系分析法（GI）；客观赋权法有：熵权法、标准差和平均差最大化法。

首先用主观赋权和客观赋权得到各自方法下的权重；其次通过矩估计解非线性规划的都最优综合得分；最后通过关联分析，得到各评价方案指标对最优方案指标的灰色关联系数，即可得到恐怖事件危害性排序并分等级，具体流程如下图。

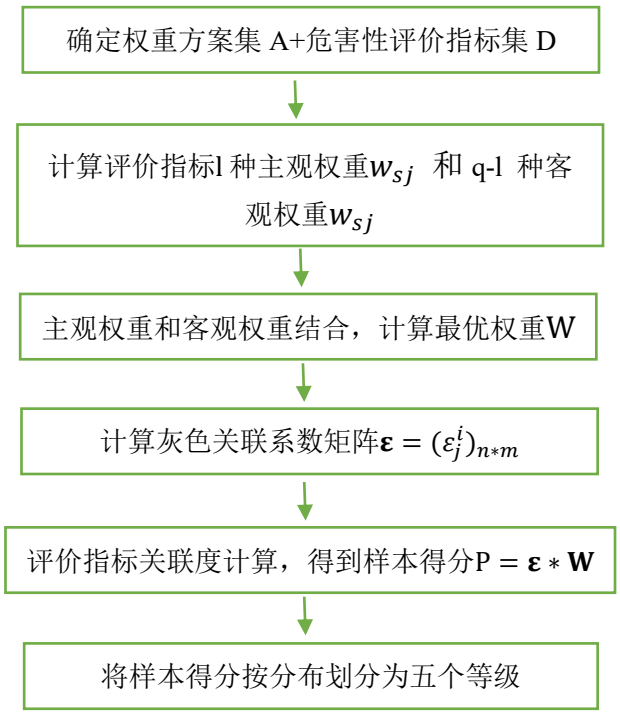

图 4.1　恐怖袭击危害程度量化分级流程图

4.1.2 最优权重计算——主客观集成法

（1）主观权重赋值法

1）改进的层次分析法（AHP）的主观权重

层次分析法的主要思想是将要研究的复杂问题看作一个受多种因素影响的大系统，进而把其中相互关联、相互制约的因素按照它们之间的隶属关系分解为若干有序层次，并根据一定客观事实对两两指标之间的重要程度作出比较判断，利用数学方法确定出表达每个层次全部元素相对重要性次序的数值，并通过对各层次的分析得出不同方案重要性程度的权重，为最佳方案的选择提供依据[1]。大体要经过以下四个步骤：

步骤 1：建立层次结构模型：目的层，准则层，方案层（因子选取在数据预处理中）。

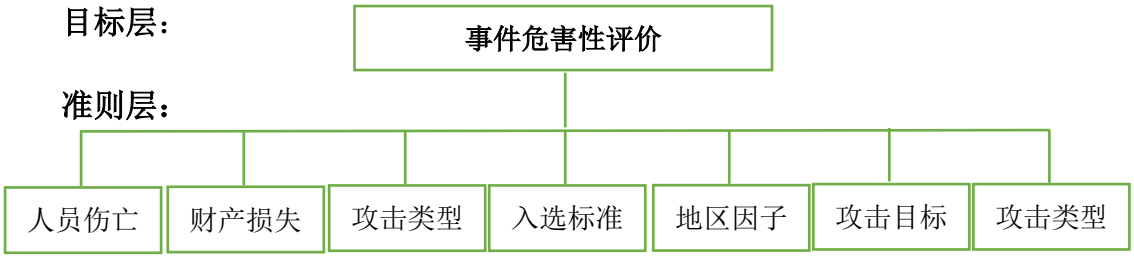

图 4.2 层次分析结构模型

表 4.1 层次分析结构：准则层-方案层

准则层	方案层
人员伤亡	伤亡因子、凶手伤亡因子
财产损失	财产损失因子
攻击类型	大规模攻击因子、绑架/挟持因子、攻击-暗杀因子、未知攻击/武器类型因子、其它武器因子、人质因子、凶手数量
入选标准	非明确入选标准因子、胁迫/恐怖入选因子
地区因子	东南亚地区因子、非洲/中东地区因子、亚洲其它地区因子、美洲地区因子、欧洲地区因子
攻击目标	城市/公共设施目标因子、商业目标因子、未知目标因子、非政府组织目标因子、宗教人物/机构目标因子、其它目标因子、目标因子-警察
国际、宗教因子	国际因子、政治/宗教入选因子

步骤 2：构造判断矩阵：相对于上一层因素重要性，采用对应 9 个评价等级标度将同一层指标的两两比较，得到判断矩阵 **A**；

$$A = (a_{ij})_{n*n} \tag{4.1}$$

式中: a_{ij}为第 i 个因素相对于第 j 个因素的比较结果。并且满足a_{ij}>0，a_{ij}=1/a_{ji}

步骤 3：层次单排序权重及一致性检验: 指同一隶属关系的各因素对隶属于

[1] 郭金玉, 张忠彬, 孙庆云. 层次分析法的研究与应用[J]. 中国安全科学学报, 2008, 18(5):148.

上一层次相应因素相对重要性的排序。相对重要性的程度用层次单排序权重来衡量。

步骤 4：一致性检验：

$$CI = \frac{\lambda_{max}-n}{n-1} \tag{4.2}$$

其中，λ_{max}为判断矩阵的最大特征值。当$CR = \frac{CI}{RI}$ =<0.10 时，认为判断矩阵的一致性是可以接受的，否则应对判断矩阵做一定修正。

2）序关系分析法(GI)的主观权重

序关系分析法通过对各危害指标按某标准排出重要性顺序来确定权重，而不用进行一致性检验。其原理及计算步骤如下[2]。

步骤 1：确定序关系，假设由 m 个危害评价指标组成的评价指标集D = $\{d_1, d_2, d_3, \dots, d_m,\}$，按照某评价准则确定了 m 个评价指标的相对重要性排序为$d_1' > d_2' > d_3' > \cdots, d_m'$。

步骤 2：确定相邻评价指标d_{j-1}的权重w_{j-1}相对于d_j的权重w_j的重要性程度之比，即权重评价标度$r_j = w_{j-1}/w_j$，$\mathrm{j} = \mathrm{m}, \mathrm{m}-1, \mathrm{m}-2$。依据评价语气算子与评价标度$r_j$ 的对应关系，确定各相邻评价指标间的权重评价标度。

步骤 3：确定评价指标的相对权重向量。

（2）客观权重赋值法

1）基于熵权的客观权重

熵权法确定客观权重的基本思想是[3]：根据指标变异性的大小来确定客观权重。一般来说，若某个指标的信息熵越小，表明指标值得变异程度越大，提供的信息量越多，在综合评价中所能起到的作用也越大，其权重也就越大。相反，某个指标的信息熵越大，表明指标值得变异程度越小，提供的信息量也越少，在综合评价中所起到的作用也越小，其权重也就越小。其计算步骤如下：

步骤 1：对标准化后的所有样本的各个危害指标计算信息熵值h_j；

$$h_j = -\frac{\log(n)}{\sum_{i=1}^{n} p_{ij}\log(p_{ij})} \tag{4.3}$$

其中，$p_{ij} = X_{ij}/\sum_{i=1}^{n} X_{ij}$，n 为样本数。

步骤 2：根据熵值，按照 h_j 大则权重系数取小、h_j 小则权重系数取大的原则确定各指标的权重系数W_{GI}。 ：

[2] 李连结，姚建刚，龙立波等．组合赋权法在电能质量模糊综合评价中的应用[J]．电力系统自动化，2007,31(4)：56-60．

[3] 聂宏展,方吕盼,乔怡等．基于熵权法的输电网规划方案模糊综合评价[J]．电网技术，2009,33(11):60-64.

$$W_j = \frac{1-h_j}{m-\sum_{j=1}^{m} h_j} \tag{4.4}$$

其中$j = 1,2,\ldots,m,$ 代表 m 个评价指标。

2）基于标准差和平均差最大化的客观权重

如果指标d_j 对有恐怖事件的属性值均无差别或差异很小，则该指标对事件危害性影响将较小，这类评价指标应给予较小的权重系数；反之，应赋予较大的权重系数[4]。可以用标准差和平均差来衡量指标d_j 属性值的差异，具体的计算步骤：

步骤 1：计算评价指标 j 对各个事件的平均值 $\overline{x_j}$ 及标准差σ_j；

步骤 2：计算评价指标 j 对各个事件的平均差u_j ：

$$u_j = \frac{1}{n}|x_{ij} - \overline{x_j}| \tag{4.5}$$

步骤 3：基于标准差和平均差最大化，对指标 j 赋予权重值：

$$w_j = \frac{z_1\sigma_j + z_1 u_j}{\sum_j^m (z_1\sigma_j + z_1 u_j)} \tag{4.6}$$

其中，z_1、z_2分别为标准差和平均差的重要性系数，$z_1 + z_2 = 1$，本文均取 0.5。

（3）主观权重与客观权重组合赋权

为了在评价体系中既反映决策的主观性，又体现决策的客观性，本章节将对主管权重和客观权重进行组合，计算最优权重。

设有 l 种主观赋权法对危害性评价指标进行赋权，则可得到按各主观赋权原则确定的各指标主观权重集合 $W_s = \{w_{sj} | 1 \le s \le l, 1 \le j \le m\}$。对 ∀s, $\sum_{j=1}^{m} w_{sj} = 1, w_{sj} \ge 0$。

采用$q - l$种客观赋权法对评价指标进行赋权，得到的客观权重集合$W_b = \{w_{bj} | l+1 \le b \le q, 1 \le j \le m\}$。对∀ b, $\sum_{j=1}^{m} w_{bj} = 1, w_{bj} \ge 0$。

假设评价指标的集成权重向量为$[w_1, w_2, \ldots, w_m]$。对于主观权重，如果决策者的数量趋于很大时，由统计学的大数定理可知其判断的权重向量集成结果将接近集成权重向量$[w_1, w_2, \ldots, w_m]$；对于客观权重，采用不同的算法得到的结果具有重复性[5]。因此，可以将其看作从总体中抽取样本来估计组合权重向量$[w_1, w_2, \ldots, w_m]$。

设分别从主观权重总体和客观权重总体中抽取 l 个样本和$q - l$个样本，针对

[4] 王应明,张军奎．基于标准差和平均差的权系数确定方法及其应用[J]．数理统计与管理，2003,22(7):22-26.

[5] 杨虎,刘琼荪,钟波．数理统计[M], 北京:高等教育出版社，2004

每个评价指标$d_j(1 \le j \le m)$，有 q 个权重样本，对于各评价指标的集成组合权重$w_j(1 \le j \le m)$，需要满足w_j与其 q 个主客观权重的偏差越小越好。同时，对于不同的评价指标，主观权重与客观权重的相对重要程度都不同，如果主观权重与客观权重的相对重要程度系数分别为 α 和 β，则集成组合权重的优化模型为：

$$\min H(w_j) = \alpha \sum_{s=1}^{l}(w_j - w_{sj})^2 + \beta \sum_{s=1}^{l}(w_j - w_{bj})^2 \tag{4.7}$$

其中，$0 \le w_j \le 1,\ 1 \le j \le m$

q 个样本分别来自 2 个总体，按照矩估计理论的基本思想，对每个评价指标$d_j(1 \le j \le m)$，其w_{sj}和w_{bj}的期望值[6]为：

$$\begin{cases} E(w_{sj}) = \frac{\sum_{s=1}^{l} w_{sj}}{l} & 1 \le j \le m \\ E(w_{bj}) = \frac{\sum_{b=1}^{q} w_{bj}}{q-l} & 1 \le j \le m \end{cases} \tag{4.8}$$

利用式(2)，可计算出每个指标$d_j(1 \le j \le m)$的主观和客观权重的重要系数α_j 和 β_j 为：

$$\begin{cases} \alpha_j = \frac{E(w_{sj})}{E(w_{sj})+E(w_{bj})} \\ \beta_j = \frac{E(w_{bj})}{E(w_{sj})+E(w_{bj})} \end{cases} \tag{4.9}$$

针对多指标决策矩阵中的评价指标，可以看成是从 2 个总体中分别取 m 个样本，同样采用矩估计理论的基本思想，可以得到:

$$\begin{cases} \alpha = \frac{\sum_{j=1}^{m} \alpha_j}{\sum_{j=1}^{m} \alpha_j + \sum_{j=1}^{m} \beta_j} = \frac{\sum_{j=1}^{m} \alpha_j}{m} \\ \beta = \frac{\sum_{j=1}^{m} \beta_j}{\sum_{j=1}^{m} \alpha_j + \sum_{j=1}^{m} \beta_j} = \frac{\sum_{j=1}^{m} \beta_j}{m} \end{cases} \tag{4.10}$$

针对每一个评价指标$d_j(1 \le j \le m)$，希望越小越好，为此，(1)式所示优化模型可以转化为：

$$\begin{cases} \min H = \{H(w_1), H(w_2), \dots, H(w_m)\} \\ \text{s.t.} \sum_{j=1}^{m} w_j = 1, \\ 0 \le w_j \le 1, 1 \le j \le m \end{cases} \tag{4.11}$$

为了求解式(5)，采用等权的线性加权方法，将多目标最优化模型转化为单目标最优化模型，即

[6] 江文奇. 多属性决策的组合赋权优化方法[J]. 运筹与管理，2006,15(6): 40-43.

$$\begin{cases} \min H = \sum_{j=1}^{m} \alpha \sum_{s=1}^{l} \left(w_j - w_{sj}\right)^2 + \sum_{j=1}^{m} \beta \sum_{b=l+1}^{l} (w_j - w_{bj})^2 \\ \text{s.t.} \sum_{j=1}^{m} w_j = 1, \\ 0 \le w_j \le 1, 1 \le j \le m \end{cases} \tag{4.12}$$

通过对式 (6) 进行求解，即可求得基于多个主客观评价指标的最优组合权向量。

4.1.3 灰色关联度分析评价

灰色理论应用最广泛的是关联分析方法，其关联系数反映各评价对象对理想（标准）对象的接近程度，关联系数越大则评价指标越优[7]。利用灰色关联分析方法求取灰色关联系数的具体步骤如下：

（1）选择参考数列。设 i 为第 i 个恐怖袭击事件，i=1,2,…,n；k 为第 k 个评价指标，k=1,2,…,n；u_{ik}为第 i 个恐怖袭击事件的第 k 个指标的取值，取各恐怖袭击事件中每个指标的最优值u_{0k}作为参考数列，于是有：

$$U_0 = (u_{01}, u_{02}, \dots, u_{0m}) \tag{4.13}$$

（2）指标值的规范化。由于评价指标间通常是有不同的量纲和数量级，因此采用极值处理法对指标值进行规范化，规范化的公式如下（本文在数据预处理阶段已进行规范化）：

$$x_{ik} = \frac{u_{ik} - min_i u_{ik}}{max_i u_{ik} - min_i u_{ik}} \tag{4.14}$$

其中 i = 1，2，…,n; k = 1,2,…,m.

（3）计算关联系数。根据灰色系统理论，将规范化后的数列X_0作为参考数列，将X_i 作为比较数列，用关联分析方法分别求得第 i 个恐怖袭击事件的第 k 个指标与第 k 个最优指标的关联系数ε_{ik}，计算公式为:

$$\varepsilon_{ik} = \frac{min_i min_k |x_{0k} - x_{ik}| + \rho * max_i max_k |x_{0k} - x_{ik}|}{|x_{0k} - x_{ik}| + \rho * max_i max_k |x_{0k} - x_{ik}|} \tag{4.15}$$

其中ρϵ[0,1]，是分辨系数，一般取 0.5，从而可以计算得到关联系数矩阵

$$\varepsilon = (\varepsilon_j^i)_{n*m} \tag{4.16}$$

根据关联系数和最优组合权重，可求得每个恐怖袭击事件关联度 P 序列，从而确定各个事件优劣排序。

$$\mathrm{P} = \varepsilon * \mathrm{W} \tag{4.17}$$

由于得分 P 为连续取值，需要对其分为五个等级。由于线性等分很多时候对于边界值的界定较为主观，因此我们将使用 Kmeans 聚类对 P 序列聚类，能够保证每一种类别得分相对集中。

[7] 刘思峰, 等. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2010

4.2 模型的求解

4.2.1 数据预处理

本题共给出了 114183 条恐怖袭击数据，共包含 134 个变量。首先对数据集进行数据清洗工作。由于附件 1 中提供了较多的变量，而一些变量明显对危害性没有任何影响，因此予以剔除；同时对缺失值超过 90%的变量也予以删除，总共保留 21 个变量。

表 4.2 变量说明

gname	犯罪集团的名称
motive	动机
individual	是否个体作案
nperps	凶手数量
claimed	声称负责
weaptype1	武器类型
nkill	死亡总数
nkillter	凶手死亡人数
nwound	受伤总数
nwoundte	凶手受伤人数
property	财产损失
propextent	财产损害程度
propvalue	财产损失的价值（美元）
ishostkid	人质或绑架的受害者
nhostkid	人质/绑架受害者总数
ransom	索要赎金
hostkidoutcome	绑架/人质结果
nreleased	释放/逃脱/获救的数量
INT_LOG	国际后勤
INT_IDEO	国际的意识形态
INT_MISC	国际杂类

（1）数值型变量处理

缺失值处理:对于数值型变量，变量缺失比例大于等于 75%则创建一个新的变量指示该变量是否缺失，即若原变量值为缺失则将新变量设为 1，否则设为 0。对于缺失比例小于 75%的变量判断为随机缺失，利用中位数进行填补。

异常值处理:将各变量低于 5%分位数或高于 95 分位数的数据分别用该变量的 5%和 95%分位数替换。

标准化处理:为了消除量纲影响对于数值型变量，利用最小-最大规范化方法进行处理。

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4.18}$$

(2) 分类型变量处理

缺失值处理：分类型变量的频数频率表见附件，将各变量中的缺失数据定义为一个新的分类。

One-hot 处理：由于数据中分类变量类别较多，比如地区变量有 12 个类别，全部做 One-hot 处理，将严重提高数据维度；因此这里将利用 word2vec 将 one-hot 矩阵降维，用更少的维度去表示地区分类。

4.2.2 因子分析特征降维

通过对附件 1 中的数据建立预处理、特征工程、特征选择，变量从原来的 21 个变成了 52 个。而特征维度太高对主观法与客观法建立评价体系来说都很复杂，因此我们利用因子分析降维，将 52 个变量降成了 26 维。这里只展示几个重要公共因子：

表 4.3 公共因子表

因子	变量 1	变量 2	变量 3
Factor1	是否涉及人质/绑架	攻击类型-劫持	是否为持续性事件
Factor2	标准 3-超出国际人道主义法律范围	目标类型-军事	疑似恐怖主义
Factor3	是否有财产损失	财产损失程度	财产损失值
Factor4	攻击类型-未知	武器-未知	武器-常规
Factor5	释放/逃脱人质数	人质数	
Factor6	攻击-轰炸/爆炸	攻击-武装袭击	自杀式
Factor8	受伤人数	死亡人数	
Factor10	凶手受伤人数	凶手死亡人数	
Factor11	标准 2：意图胁迫、恐吓或煽动更多的群众	暴力/恐怖	
Factor12	警察	平民	
Factor15	攻击-暗杀	是否成功	攻击-轰炸/爆炸

降维后将得到旋转后的因子载荷矩阵 A，旋转过后的因子也具有解释性：Factor1 可以看成是绑架型因子，Factor2 非明确恐怖主义因子，Factor3 财产因子，Factor4 常规攻击类型因子；Factor5 人质因子， Factor6 大规模攻击因子，Factor8 伤亡因子，Factor10 凶手伤亡因子；Factor11 暴力恐怖因子。

将因子载荷矩阵 A 与原数据矩阵相乘，得到每个样本在每个因子下的得分 B，本章节后续建模都将对因子得分 B 进行操作。

4.2.3 最优权重计算

利用 python,我们将上述主观权重和客观权重计算实现，得到结果如下：

改进的 AHP 法：建立恐怖袭击事件一级评价指标的判断矩阵 K，判断矩阵 K 的一致性比例 CR＝0.012＜0.1，满足一致性要求。求最大特征值对应的特征向量并进行归一化处理，得到一级指标的主观权重向量：

$$W = [0.28 \quad 0.220 \quad 0.10 \quad 0.04 \quad 0.07 \quad 0.14 \quad 0.16]$$

同理，通过建立二级评价指标的判断矩阵，并最终得到表的权重和矩阵。结果显示，财产损失因子$w_3 = 0.22$，伤亡因子$w_8 = 0.21$，大规模攻击因子$w_6 = 0.19$。

GI 序分析法：各因子间相对重要性程度为：伤亡因子$w_8 = 0.26$、大规模攻

击因子$w_6 = 0.21$、财产损失因子$w_3 = 0.19$、绑架/挟持因子$w_1 = 0.14$,从而得到权重系数如表 4.4。

熵权法：熵权法可以根据实际数据得到客观权重，反应实际中，某些数据可能聚集性较强，从而熵越小，权重也会相对比较小。权重较大的几个因子为伤亡因子w_8 =0.44、财产损失因子w_3 =0.30、未知目标因子w_9 =0.13。

标准差和平均值法：该方法计算出来的各个因子权重差别不是特别大，其中最重要的因子使非洲/中东地区因子w_{13} =0.13，而实际上非洲/中东地区的恐怖袭击事件也较为频繁，由于人口密集且经济较为落后，较容易造成相对危害性比较严重的恐袭事件。

表 4.4 各指标权重计算

Factor	因子含义	主观赋权法		客观赋权法		最优权重 W
		改进的 AHP 法	GI 法	熵权法	标准差和平均值法	
Factor8	伤亡因子	0.21008	0.26261	0.44198	0.02472	0.262631
Factor3	财产损失因子	0.22125	0.18887	0.30359	0.00529	0.151264
Factor6	大规模攻击因子	0.18761	0.2133	0.00504	0.07258	0.11034
Factor1	绑架/挟持因子	0.11573	0.1422	0.00058	0.02225	0.065028
Factor9	未知目标因子	0.00219	0.00393	0.13061	0.035	0.05047
Factor2	非明确入选标准因子	0.07134	0.06348	0.00433	0.05264	0.042831
Factor13	非洲/中东地区因子	0.02534	0.01306	0.00976	0.11365	0.041762
Factor15	攻击-暗杀因子	0.01892	0.01005	0.04393	0.08303	0.041574
Factor18	东南亚地区因子	0.05782	0.04232	0.00153	0.04147	0.030034
Factor11	胁迫/恐怖入选因子	0.02003	0.01959	0.00739	0.06321	0.028862
Factor23	亚洲其它地区因子	0.00015	0.00037	0.0213	0.06256	0.024894
Factor19	城市/公共设施目标因子	0.03962	0.02351	0.00113	0.02716	0.017984
Factor4	未知攻击/武器类型因子	0.00143	0.0008	0.00393	0.05375	0.017346
Factor12	目标因子-警察	0.0006	0.0002	0.00259	0.04384	0.013779
Factor17	商业目标因子	0.00923	0.00628	0.00094	0.03544	0.013307
Factor14	欧洲地区因子	0.00012	0.00028	0.00281	0.03535	0.011377
Factor21	非政府组织目标因子	0.00286	0.00128	0.00159	0.0342	0.011082
Factor7	国际因子	0.00167	0.00231	0.00065	0.03324	0.010944
Factor20	政治/宗教入选因子	0.00245	0.00192	0.00091	0.03326	0.010871
Factor16	美洲地区因子	0.00056	0.00031	0.00101	0.03297	0.010153
Factor22	宗教人物/机构目标因子	0.0012	0.0008	0.00068	0.03085	0.009631
Factor24	其它武器因子	0.00032	0.00053	0.00173	0.02568	0.008305

Factor26	其它目标因子	0.00041	0.00049	0.00044	0.02095	0.006512
Factor25	凶手数量因子	0.0005	0.0008	0.00837	0.00726	0.00494
Factor10	凶手伤亡因子	0.00089	0.00053	0.00144	0.00616	0.00246
Factor5	人质因子	0.00004	0.00016	0.00177	0.0035	0.001617

最优权重组合：根据模型公式可计算出基于上述主观权重和客观权重的重要程度系数为$\alpha = 0.5815$，$\beta = 0.4185$.在此基础上，通过对非线性规划求解，可以得到评价指标权重的最优组合（见表 4.4）。其中，伤亡因子成为权重最大的因子$w_8 = 0.26$，其次是财产因子$w_3 = 0.15$，大规模攻击因子$w_6 = 0.11$；相对来说，凶手数量因子$w_{25} = 0.00494$、凶手伤亡因子$w_{10} = 0.00246$、人质因子$w_5 = 0.00162$ 权重较低。

对比上述权重向量可知，采用主观赋权和客观赋权得到的权重赋值相互差别较大，而基于矩估计理论最优组合赋权得到的各赋权值之间相互差别较小，赋权结果更趋合理。

4.2.3 灰色综合评价划分危害等级

对于恐袭危害性评价指标，以最理想的参数为参考序列，取各评价指标的理想参数值为 0，则理想样本的评价指标最优值 $\boldsymbol{R}^* = \mathbf{0}$。通过建立评价指标和样本组成的数据矩阵，经无量纲化之后，形成数据决策矩阵 **R**，将评价指标样本数据与理想参数值进行比较，形成差值矩阵，可计算各评价指标与各等级的关联系数矩阵$\boldsymbol{\varepsilon}$。

同时，计算各样本与理想样本之间的关联度$\mathbf{P} = \boldsymbol{\varepsilon} * \mathbf{W}$ ，由于线性等分很多时候对于边界值的界定较为主观，因此我们将得到的关联度 **P** 序列按聚集性程度聚成 5 类，每种类别相对距离比较集中，危害性逐级递增划分结果如下：

表 4.5 等级划分结果

等级	样本数	综合得分均值	综合得分范围
1	3	25.12	(11.43,+∞)
2	38	3.41	(1.99, 11.43)
3	12848	0.52	(0.43,1.99)
4	34417	0.34	(0.105,0.43)
5	66877	-0.13	(-∞,0.105)

等级划分结果显示，危害等级 1（危害性最强）的事件有 3 起，且综合得分范围(11.43,+∞)；等级 2 的事件有 38 起，综合得分范围(1.99, 11.43)；等级 3 的事件有 12848 起，综合得分范围(0.43,1.99)；等级 4 的事件有 34417 起，综合得分范围(0.105,0.43)；等级 5 的事件有 66877 起，综合得分范围(-∞,0.105)。

表 4.6 危害前十的恐怖袭击事件

事件编号	综合得分	等级
200109110005	28.352	1
200109110004	28.347	1
200107240001	18.667	1
199808070002	11.431	2
201406150063	8.936	2
201602180049	5.357	2
201412070129	5.318	2
201710140002	4.878	2
201406100042	4.314	2
201408200027	4.135	2

从危害前十的事件中，等级为 1 的恐袭事件分别为：200109110005、200109110004、200107240001，前两者均为死伤非常严重的 911 事件，民航客机撞上纽约世贸双子星大楼，最终造成近 3000 人死亡，2 座建筑物倒塌，7 座建筑物局部毁损并坍塌，也被公认为目前恐怖袭击事件中最严重的一次事件。

表 4.7 典型事件危害级别

事件编号	危害级别
200108110012	2
200511180002	3
200901170021	2
201402110015	3
201405010071	3
201411070002	4
201412160041	3
201508010015	5
201705080012	5

五、 问题二的建模与求解

5.1 模型的构建

本文对问题二的模型构建可以划分为三个步骤，如图 5.1 所示。第一步是基于 OPTICS 算法对恐袭事件进行聚类分析。对 2015、2016 年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件进行聚类，使得同一类别中的恐怖袭击事件的特征尽可能相似，不同类别之间的恐怖袭击事件的特征尽可能不同，并认为同一类别内的恐怖袭击事件是由同一恐怖组织或个人发起的，而不同类别的恐怖事件则不是。第二步是从恐怖袭击发起数量和危害程度（利用问题一结果）两个方面进行考量，对恐怖组织或个人的危害性进行综合评分，通过对得分进行排序找出危害性最大的 5 个恐怖组织或个人。第三步是利用 XGBoost 算法对题目中所给典型事件按嫌疑程度给第二步中得到的 5 个恐怖组织或个人进行排序。

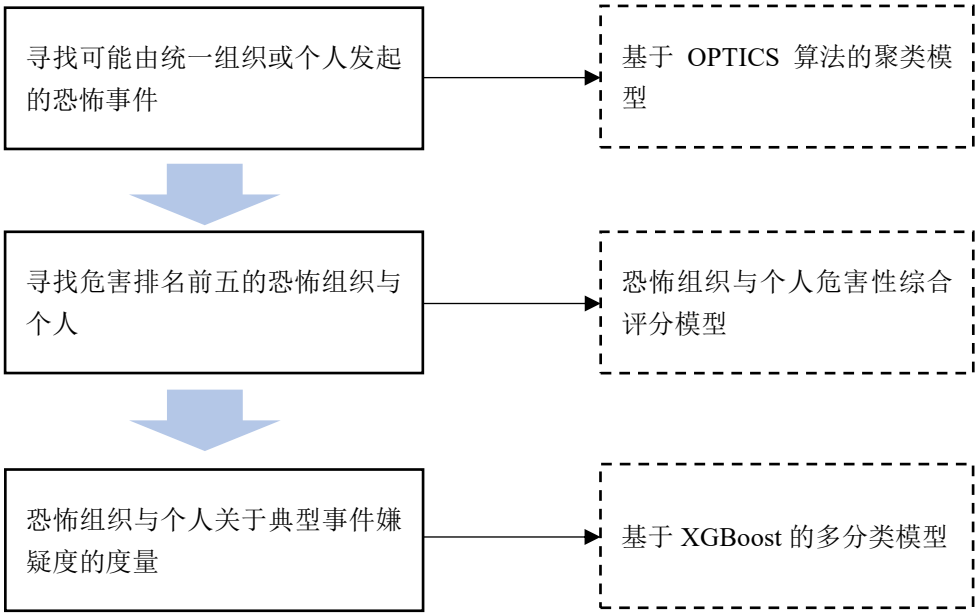

图 5.1 问题二求解过程及方法

5.1.1 基于 OPSTIC 算法的恐怖袭击事件聚类模型

1）DBSCAN 算法

基于密度的聚类算法是数据挖掘技术中被广泛应用的一类算法，其核心思想是用一个点的ε邻域内的邻居点数来衡量该点所在空间的密度。应用这种算法可以找出形状不规则的类，且在聚类前无需指定类的个数。

DBSCAN[8](Density-Based Spatial Clustering of Applications with Noise)就是一个比较有代表性的基于密度的聚类算法。DBSCN 算法中有两个重要参数，分别

[8] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996:226-231.

为定义密度时的领域半径ϵ和定义核心点时的阈值 M。为更好地阐述该算法的基本思想，现考虑数据集合 X=$\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$并引入以下概念与记号。

(1) ϵ邻域(ϵ neighborhood):

设 x∈X，定义$N_\epsilon(x)$为 x 的ϵ邻域

$$N_\epsilon(x) = \{y \in X:\ d(y,x) \le \epsilon\} \tag{5.1}$$

有时为简单起见也将节点$x^{(i)}$与其指标 i 视为等同（因为它们一一对应），并引入记号:

$$N_\epsilon(i) = \{j: d(x^{(j)}, x^{(i)}) \le \varepsilon,\ x^{(j)}, x^{(i)} \in X\} \tag{5.2}$$

(2) 密度(density):

设 x∈X，定义 x 的密度为：

$$\rho(x) = |N_\varepsilon(x)| \tag{5.3}$$

这里的密度是一个整数值，且依赖于半径ϵ。

(3) 核心点(core point):

设 x∈X，若 ρ(x)≥M，则称 x 为 X 的核心点。记由 X 中所有核心点构成的集合为X_c，并令N_{nc}=X\X_c，表示由 X 中的所有非核心点构成的集合。

(4) 边界点(border point):

若 x∈X_{nc}，且存在 y∈X，满足

$$y \in N_\epsilon(x) \cap X_c \tag{5.4}$$

即 x 的ϵ邻域中存在核心点则称 x 为 X 的边界点,且记X_{bd}为由 X 中所有边界点构成的集合。

(5) 噪音点(noise point):

记$x_{noi} = X/(X_c \cup X_{bd})$，若 x∈$X_{noi}$，则称 x 为噪音点。

图 5.2 可以帮助更清晰直观的区分核心点、边界点和噪音点

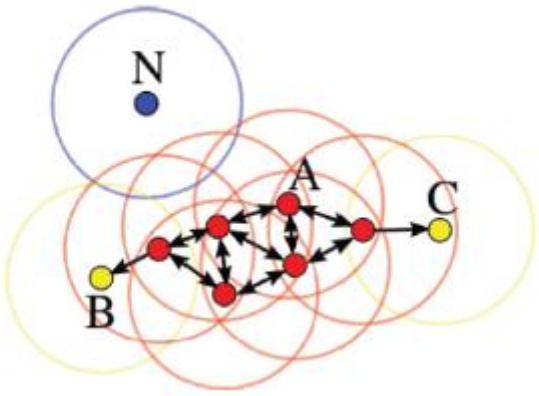

图 5.2 核心点、边界点和噪音点（红黄蓝）示意图

(6) 直接密度可达(directly density-reachable):

设 x,y∈X，若 x,y 同时满足 x∈X_c，y∈N_ϵ(x)，则称 y 是从 x 直接密度可达的。

(7) 密度可达(density-reachable):

设$p^{(1)}, p^{(2)}, \ldots p^{(m)}$ ∈ X，且 m≥2。若$p^{(i+1)}$是从$p^{(i)}$直接密度可达的，i=1,2…,m-1，则称$p^{(m)}$是从$p^{(1)}$密度可达的。

(8) 密度相连(density-connected):

设 x,y,z∈X，若 y 和 z 均是从 x 密度可达的，则称 y 和 z 是密度相连的。

(9) 类(cluster):

称非空集合 C ⊆X 是 X 的一个类，若对于 x，y∈X 满足以下条件：

a) 若 x∈C，且 y 是从 x 密度可达的，则 y∈C

b) 若 x∈C，y∈C，则 x，y 密度相连

基于以上概念，可将 DBSCAN 算法的核心思想概括为为：从某个选定的核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

2) OPTICS 算法

上文介绍的 DBSCAN 算法虽然具有速度快、能够处理噪声点、聚类簇的形状没有偏倚、无需设置聚类个数等种种优点，但是它的对于初始参数领域半径ϵ和阈值 M 的取值非常敏感。由于题目所给恐怖袭击数据库维度较多、数据结构较为复杂，在使用 DBSCAN 算法时难以确定比较合适的参数。为克服 DBSCAN 算法的这一缺点，Ankerst Breunig 和 Kriegel 提出了 OPTICS[9]算法，OPTICS 算法并不直接寻找各个簇, 而是将基于密度查找簇所需要的信息记录下来, 这些信息反映了数据空间基于密度的簇结构。同时, 从这些密度信息也可以直接发现各个簇。OPTICS 在 DBSCAN 算法的基础上引入了如下两个新的概念：

(1) 核心距离(core-distance)

设 x∈X，对给定的参数ϵ和 M，称使得 x 成为核心点的最小邻域半径为 x 的核心距离，即：

$$\text{cd(x)}=\begin{cases} UNDEFINED, & \text{若}|N_\varepsilon(\text{x})| < \text{M} \\ d\left(\text{x}, N_\epsilon^M(\text{x})\right), & \text{若}|N_\varepsilon(\text{x})| \geq \text{M} \end{cases} \tag{5.5}$$

其中，N_ϵ^i(x)表示集合N_ε(x)中与节点 x 第 i 近邻的节点。

(2) 可达距离(reachability-distance)

设 x，y∈X，对于给定参数ϵ和 M，将 y 关于 x 的可达距离定义为：

[9] Ankerst M, Breunig M M, Kriegel H P. OPTICS:ordering points to identify the clustering structure[J]. Acm Sigmod Record, 1999, 28(2):49-60.

$$\mathrm{cd(x)}=\begin{cases} UNDEFINED, & 若|N_\varepsilon(\mathrm{x})|<\mathrm{M} \\ \max\{\mathrm{cd(x)},\mathrm{d(x,y)}\}, & 若|N_\varepsilon(\mathrm{x})|\geq\mathrm{M} \end{cases} \tag{5.6}$$

5.1.2 恐怖组织及个人危害性综合评分模型

本文从恐怖组织及个人发动的恐怖袭击数量及危害性两个角度考虑定义了一个恐怖组织危害性综合评分模型。模型的符号和公式如下：

(1) i 表示聚类得到的第 i 个恐怖组织或个人，i=1,2…n;

(2) n_{ij}表示由第 i 个恐怖组织或个人发起的危害程度为 j 级别的恐怖袭击事件数，j=1,2…5;

(3) $Score_i$表示第 i 个恐怖组织或个人的危害程度得分

$$Score_i = \sum_{j=1}^{5} n_{ij} * (6-j) \qquad \text{i=1,2…n} \tag{5.7}$$

对公式(5.7)计算的值按从大到小进行排序，排名前 5 的即为前五大恐怖组织或个人。

5.1.3 基于 XGBoost 的恐怖组织或个人嫌疑判断模型

XGBoost[10]（Extreme Gradient Boosting）算法是一种基于梯度提升算法（Gradient Boosting）以及决策树（DecisionTree）的改进型学习算法。其原理是使用迭代运算的思想，将大量的弱分类器转化成强分类器，以实现准确的分类效果。

XGBoost 是 Boosting 中的经典方法，它与决策树是息息相关的，它通过将很多的决策树集成起来得到一个很强的分类器。其中决策树就是一种对空间不断进行划分算法，通过给每个划分的空间赋予一个标签或权重，那么当样本落到这个空间里面，就认为这个样本就满足这个标签。Boosting 的核心思想就是希望训练出 K 颗树，并将它们集成起来用于预测 Y。

XGBoost 算法的原理是将原始数据集分割成多个子数据集，将每个子数据集随机的分配给基分类器进行预测，然后将弱分类的结果按照一定的权重进行计算，一次来预测最后的结果。一般将这种迭加基分类器结果的预测模型称为加性模型。公式如下：

$$\mathrm{F}=\sum_{i=1}^{m-1} f + f_m \tag{5.8}$$

加性模型的预测拟合是对每个基分类器的预测结果进行向后拟合算法的样条平滑，使得拟合误差即偏度和自由度即方差之间达到均衡状态。本文基函数选择的是回归树，有的文献也将其称为 CART 树，选择回归树是立足于我们原有的经验。当树的深度足够深时，其作为分类器的效果会非常的好。因此，选择 CART 树作为模型的基函数，那么单个 CART 第 M 次预测的结果为：

$$f_m(x) = T(X;\theta_m) \tag{5.9}$$

由此，基函数已经确定，其中 T 为决策树，m 代表基分类器的数量，theta

[10] Chen T, Guestrin C. XGBoost:A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.

代表决策树的划分路径，最后的预测结果为前一次的预测结果加上当下的决策树，而误差项可以表示为：

$$L(y,\hat{y})=L(y,f_{m-1}(x))+T(X,\theta_m) \tag{5.10}$$

这里 $L(y, \hat{y}_i)$是真实值y_i和预测值$\hat{y}_i$之间差值之和。

基尼系数、剪枝和控制树的深度是 CART 进行分类的重要手段。实际是通过上述方式控制模型的方差和偏差，使得模型的拟合泛化能力更强。例如，可以用叶子节点数目 T 和 Leaf Score 的 L2 模的平方来定义结构复杂度函数：

$$\emptyset(\theta)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\|W\|_j^2 \tag{5.11}$$

其中γ表示控制树复杂度的系数，相当于给 XGBoost 算法模型的树做了前剪枝，而λ表示通过多大的比例来改变正则项，相当于给复杂的模型一个惩罚，防止模型出现过拟合。在综合偏差函数和方差函数后给出以下目标函数：

$$Obj(\theta)=\sum_i l(y_i+\hat{y}_i)+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\|W\|_j^2 \tag{5.12}$$

之前学者对于 Gradient Boosting Descent Tree 算法的处理方法是反复计算目标函数的误差，使其误差越来越小，这种方法即为梯度下降算法。只要每次都按照这种方式得出弱分类器，再将每个弱分类器进行相加，最后的模型得出的结果必然是最优的。梯度下降的公式如下所示：

$$-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]f(x)=f_{m-1}(x) \tag{5.13}$$

按上述公式对损失函数求导之后，树的个子节点和每个节点的权重都可以确定，所以树也就确定了。但是，由于 XGBoost 算法所使用的基分类器数量较多，因此，需要运用更为通用的算法来实现梯度下降，XGBoost 算法使用了泰勒二阶展开替代了原来的一阶导数，使得算法更具普遍性。加入泰勒二阶展开之后的目标函数如下：

$$Obj(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{m-1}+f_m(x_i))+\emptyset(f_m) \tag{5.14}$$

其中，n 表示样本数量，m 表示当前迭代的次数，f(m)表示当前迭代的误差。

用泰勒展开：

$$F(x+\Delta x)\cong f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2 \tag{5.15}$$

定义：

$$g_i=\frac{\partial l(y_i,\hat{y}_i^{m-1})}{\partial \hat{y}_i^{m-1}} \qquad h_i=\frac{\partial^2 l(y_i,\hat{y}_i^{m-1})}{\partial \hat{y}_i^{m-1}} \tag{5.16}$$

代入计算：

$$Obj_m\cong\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{m-1})+g_i f_m(x_i)+\frac{1}{2}h_i f_m^2(x_i)\right]+\emptyset(f_m) \tag{5.17}$$

5.2 模型的求解

5.2.1 数据预处理

从样本中找出 2015、2016 年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件共 22743 起。根据变量的实际意义构造特征向量，本题中新构造的变量名称及含义如下表（全部变量表见附件）：

表 5.1 新构造特征向量汇总表

变量名	类型	含义
imonth	分类	恐袭事件发生月份
iweekday	分类	恐袭事件发生星期数
gname1	分类	是否为发起恐袭事件数量排名第一的组织
gname2	分类	是否为发起恐袭事件数量排名第二的组织
gname3	分类	是否为发起恐袭事件数量排名第三的组织
gname4	分类	是否为发起恐袭事件数量排名第四的组织
gname5	分类	是否为发起恐袭事件数量排名第五的组织

5.2.2 基于 OPSTICS 算法的恐怖袭击事件聚类模型求解

利用 OPTICS 算法对 2015、2016 年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件进行聚类，最终样本被分为了 28 个类（由于篇幅所限，聚类结果详见附件）。

5.2.3 恐怖组织危害性排序模型求解

按照公式(5.7)计算得到各恐怖组织或个人发动恐怖袭击危害性的综合得分，按照从大到小的顺序进行排序，得到危害性最大的 5 个恐怖组织或个人，这 5 个类别内的恐怖袭击事件数目如表 5.2。可以看出，各类别的样本数量比较接近，因此后面应用 XGBoost 算法建立分类模型是适宜的。

表 5.2 前五大恐怖组织或个人类别内数目

编号	疑似由该组织发动的恐怖袭击数量
1	1774
2	1296
3	1284
4	1245
5	2581

5.2.4 基于 XGBoost 的恐怖组织或个人嫌疑判断模型求解

1）样本划分

按照 8:2 的比例将 2015、2016 年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件随机划分为训练集和验证集，得到训练集中样本数为 18194 条，验证集为 4549 条。

2）模型构建与检验

对训练集中样本构建 XGBoost 模型，由于本文需要建立一个多分类模型因此将目标设置为 softprob。运用贪心算法对 XGBoost 中模型中参数进行调整，实验结果表明将参数设置为 min_child_weight=6，max_depth=7， gamma=0.025，lambda=1，Subsample=0.7，Colsample_bytree=0.6，eta=0.1，max_delta_step=0.5 时模型的效果最好。运用测试集数据对模型效果进行检验，模型依然具有良好表现。

3）典型事件划分

将题目所给典型事件带入上文训练好的 XGBoost 模型中，得到的结果表如下：

表 5.3 恐怖分子关于典型事件的嫌疑度

事件编号	1 号嫌疑人	2 号嫌疑人	3 号嫌疑人	4 号嫌疑人	5 号嫌疑人
201701090031	1	5	3	4	2
201702210037	5	1	4	3	2
201703120023	4	1	5	2	3
201705050009	4	5	1	2	3
201705050010	4	5	1	2	3
201707010028	5	4	3	2	1
201707020006	4	3	5	2	1
201708110018	3	4	5	2	1
201711010006	2	1	3	5	4
201712010003	2	4	5	3	1

六、问题三的建模与求解

从原始数据中提取 2015、2016 和 2017 年发生的 39452 起恐怖袭击事件，重点研究了近年来恐怖袭击发生的时空分布总体特征、蔓延趋势、空间聚集性等方面，并根据研究结果对未来恐怖主义发展态势及反恐工作的开展给出了预测与建议。问题三的研究思路如下图：

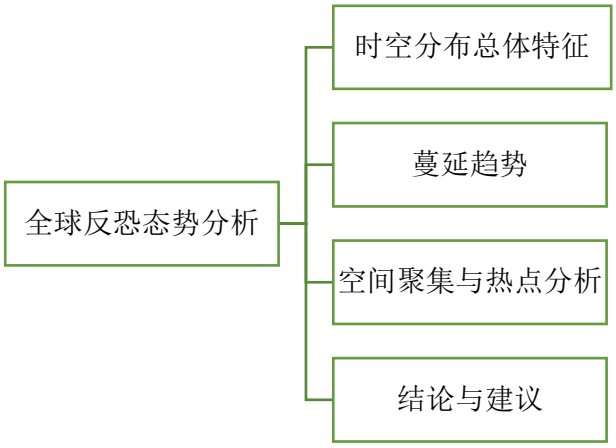

图 6.1 问题三研究思路框架图

6.1 全球恐怖袭击时空分布总体特征

研究全球恐怖袭击事件在时空上的演变有助于各国清晰把握全球恐怖主义发展趋势，并为反恐计划的制定提供重要参考，接下来本文将分别从时间和空间角度进行分析探讨。

从时间上看，近三年来全球恐怖袭击事件数量和成功率均呈现持续下降趋势，从侧面反映了近年来全球反恐工作取得了一定成效。

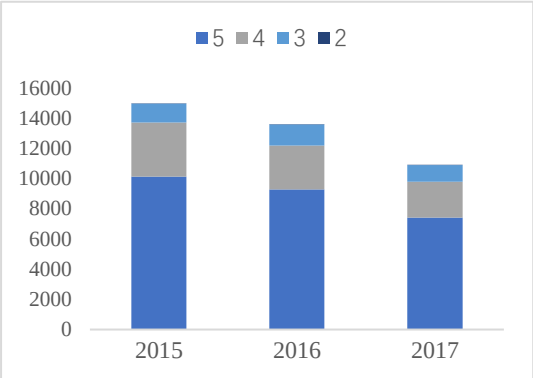

图 6.2 近三年恐怖袭击数量级别趋势

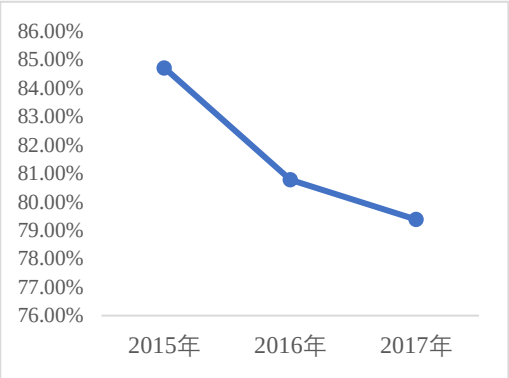

图 6.3 近三年恐怖袭击成功率趋势

从地域上看，恐怖袭击事件呈现出很强的聚集特点。过去三年，全球恐怖袭击事件发生地集中在中东及北非、南亚和撒哈拉以南的非洲地区，在这三个地区发生的恐怖袭击数在总数量中的占比分别为 40%、30%和 15%，累计占比超过 85%。此外，各地区发生恐袭事件的危害性分布也明显不同。在撒哈拉以南的非

洲、北美以及中亚地区发生危害等级较高的恐袭事件的概率明显高于其他地区。

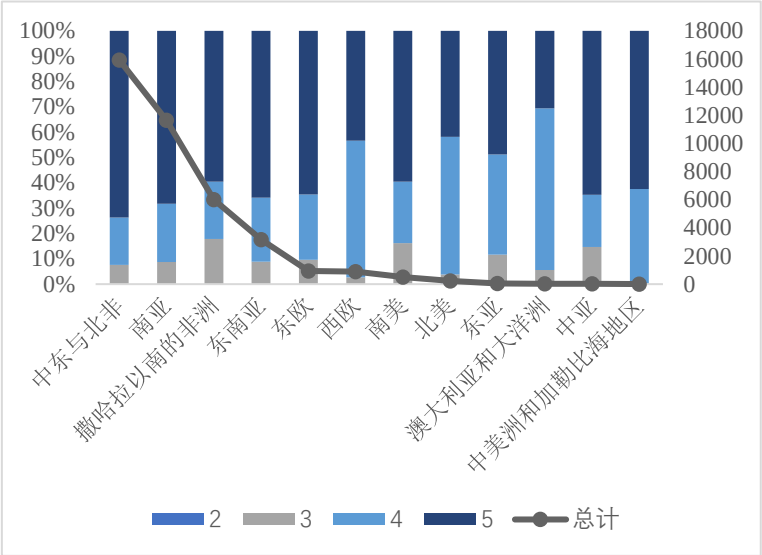

图 6.4 近三年各地区恐怖袭击数量级别分布图

6.2 恐怖主义蔓延趋势

近三年来，恐怖组织在全球的活动空间总体受到一定打压，但个别区域恐怖主义有进一步蔓延风险。2015 年全球有 99 个国家遭受了恐怖袭击，2016 年这一数字上升到 108，2017 年则下降至 102。从地区角度看，近三年来 12 个地区中只有 2 个地区的恐怖袭击事件数量持续增长，5 个地区的恐怖袭击数量则持续下降，剩余 5 区域数量在不同年度数量增减不一。其中东欧恐怖袭击数量减少幅度最大为 83%，其次是东亚地区为 75%，反映了这两个地区近年来在反恐方面取得了很大成效。但值得注意的是，中美和加勒比海地区恐怖主义开始逐渐出现，北美地区恐怖主义则有进一步扩散的风险(可利用图 6.5 到 6.7 进行对比，不同颜色代表不同等级的恐怖袭击事件)。

表 6.1 2015-2017 年各地区恐怖袭击事件数量及变化

地区	2015 年数量	2016 年数量	2017 年数量	近 2 年增长率
东欧	684	134	110	-83.92%
东亚	28	8	7	-75.00%
中东和北非	6036	6115	3780	-37.38%
中亚	10	17	7	-30.00%
南亚	4585	3639	3430	-25.19%
澳大利亚和大洋洲	14	10	12	-14.29%
西欧	333	273	291	-12.61%
东南亚	1072	1077	1020	-4.85%
南美	176	159	172	-2.27%
撒哈拉以南的非洲地区	1964	2077	1970	0.31%
北美	62	75	97	56.45%
中美和加勒比海地区	1	3	4	300.00%
总计	**14965**	**13587**	**10900**	**-27.16%**

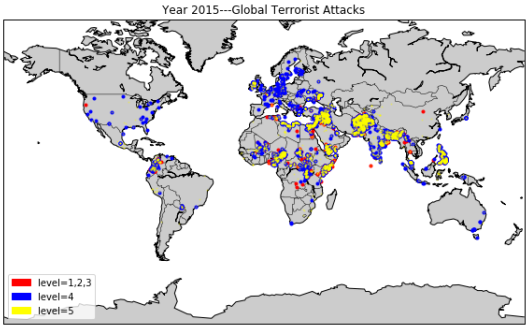

图 6.5 2015 年全球恐怖袭击分布图

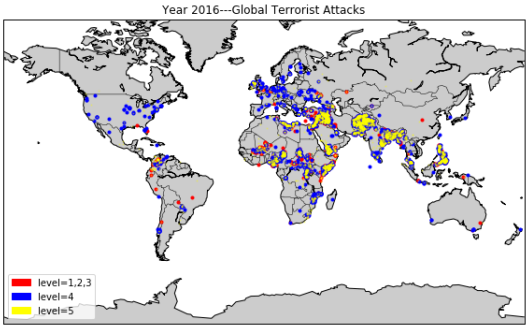

图 6.6 2016 年全球恐怖袭击分布图

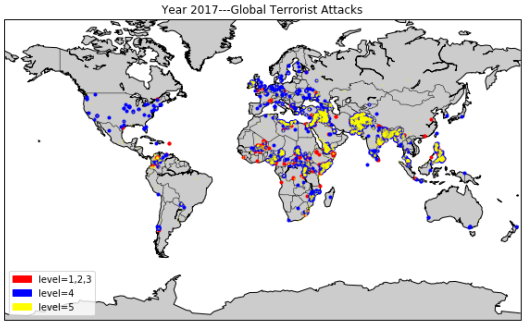

图 6.7 2017 年全球恐怖袭击分布图

6.3 空间聚集与热点分析

6.3.1 空间集聚性模型构建

(1)Lorenz 曲线与基尼系数

Lorenz 曲线与基尼系数最初被用于研究收入分配差距问题。Lorenz 曲线的构造方法如图 6.8 所示，先画一个矩形，矩形的高衡量社会财富的百分比，将之分成五等份，每一等份代表 20%的社会总财富。在矩形的宽上，将不同家庭从最贫者到最富者自左向右排列，也分为 5 等份，第一个等份代表收入最低的 20%的家庭。在这个矩形中，将每一份的家庭所有拥有的财富的百分比累计起来，并将相应的点画在图中，便得到了一条曲线就是洛伦兹曲线。

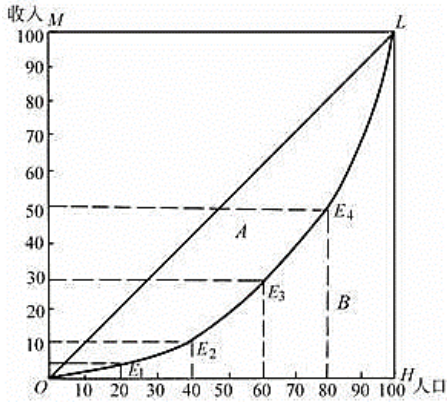

图 6.8 Lorenz 曲线示例

赫希曼根据洛伦茨曲线提出判断分配平等程度的指标基尼系数。设实际收入分配曲线和收入分配绝对平等曲线之间的面积为 A，实际收入分配曲线右下方的面积为 B。则

$$Gini=A/(A+B) \tag{6.1}$$

本文参考李国辉[11]的研究方法将其应用于对恐怖袭击的研究。在构造 Lorenz 曲线时首先把不同因素所对应的恐怖袭击频率由低到高进行排序，将各因素对应事件的累计百分比作为 y 轴，x 轴则为每个因素占全部因素个数的累计百分比。若恐怖袭击事件在某变量不同类别上呈均匀分布，即恐怖袭击在该变量上不存在热点，则 Lorenz 曲线对应的方程应该为 y=x。

(2) 帕累托分析法

帕累托分析法是由意大利经济学家维尔弗雷多·帕累托首创的常用于项目管理的一种方法。主要原理是依据事物在技术或经济方面的主要特征，进行分类排队用以区分事物的主要因素、次要因素和一般影响因素，从而有区别地确定管理方式。由于它把被分析的对象分成 A、B、C 三类，所以又称为 ABC 分析法。A 类因素是主要影响因素累计发生频率为 0%~70%。B 类因素是次要影响因素累计发生频率为 70%~90%。C 类因素是一般影响因素累计发生频率为 90%~100%，本文选取主要影响因素作为热点。

6.3.2 恐怖袭击事件特征的空间集聚性与热点

(1) 恐怖袭击事件不同因素的空间集聚性与热点

基于上文分析结果，笔者选取了在考察恐怖袭击事件特征时比较重要的 5 个分类变量进行空间集聚性探究与热点分析。首先，画出各变量的 Lorenz 曲线并计算基尼系数。由于篇幅限制，本文仅以地区（Region）变量为例对求解过程进行展示。地区变量共有 12 个分类，得到 Lorenz 曲线如图 6.9，基尼系数为 0.7175。

表 6.2 基尼系数计算过程示例

地区	事件频数	事件频率	事件累计频率	类别累计频率
中美洲和加勒比海地区	8	0.02%	0.02%	8.33%
中亚	34	0.09%	0.11%	16.67%
澳大利亚和大洋洲	36	0.09%	0.20%	25.00%
东亚	43	0.11%	0.31%	33.33%
北美	234	0.59%	0.90%	41.67%
南美	507	1.29%	2.18%	50.00%
西欧	897	2.27%	4.46%	58.33%
东欧	928	2.35%	6.81%	66.67%
东南亚	3169	8.03%	14.84%	75.00%
撒哈拉以南的非洲	6011	15.24%	30.08%	83.33%
南亚	11654	29.54%	59.62%	91.67%
中东和北非	15931	40.38%	100.00%	100.00%
总计	**39452**	**100.00%**	-	-

[11]李国辉. 全球恐怖袭击时空演变及风险分析研究[D]. 中国科学技术大学, 2014.

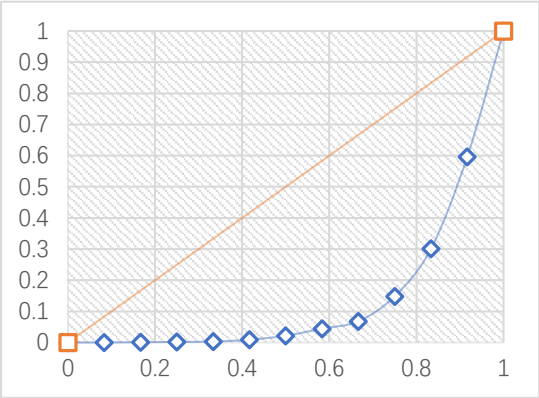

图 6.9 地区变量 Lorenz 曲线图

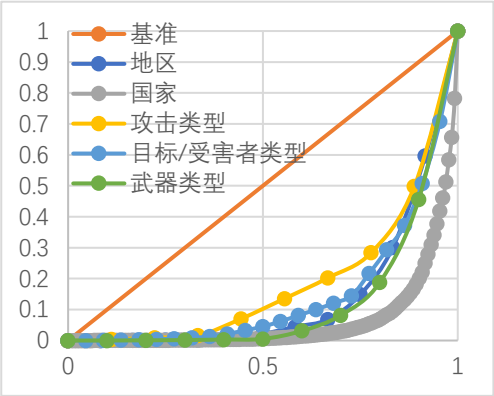

图 6.10 各变量 Lorenz 曲线图

图 6.10 为全部 5 个变量的 Lorenz 曲线，各变量基尼系数见表 6.3。基尼系数越大说明分布越不均衡即聚集性越明显。由该表可以看出 5 个变量的基尼系数均大于 0.6，即全部具有明显的聚集现象。其中，国家因素的基尼系数最高为 0.87，而地区因素的基尼系数也达到了 0.72，因此恐怖袭击的地域聚集性十分明显，即恐袭事件大多集中发生在少数国家或地区。

表 6.3 各变量的基尼系数

变量名称	基尼系数	热点
国家	0.8705	伊拉克、阿富汗、印度、巴基斯坦、菲律宾、尼日利亚、索马里、也门、埃及、叙利亚、利比亚、土耳其、泰国、乌克兰
武器类型	0.746345	爆炸物、轻武器
地区	0.717454	中东与北非、南亚、撒哈拉以南的非洲
目标/受害者类型	0.705094	公民自身和私有财产、军事、警察、政府（一般意义）、商业
攻击类型	0.617961	轰炸/爆炸、武装袭击、劫持人质(绑架)

(2) 恐怖袭击事件的空间聚集性和热点

由上文分析可以看出，恐怖袭击在全球各地区分布极不平衡，为了进一步对全球恐怖袭击事件的特点进行挖掘，为未来反恐工作的开展提供更具有参考意义的支持。本文选择以国家或地区为单位进行进一步的研究。由于国家变量有多达 132 个分类，且部分类别包含的事件数过少，因此本文选择地区变量作为划分依据。接下来将从武器类型、目标或受害者类型、攻击类型三个角度分别进行研究。

1) 武器类型的空间聚集性

近三年在全球恐怖袭击事件中的一般武器类型可划分为 9 类，由于假武器类别的事件仅为一起，故不再作考虑。在剩余的 8 个类别中，基尼系数最大的是交通工具为 0.84。使用交通工具进行恐袭的热点地区为中东与北非以及西欧，这可能是因为中东与北非原本就是恐怖袭击多发地，而西欧由于近年来难民的涌入以及恐怖组织通过网络散布极端思想更为便利，“独狼式”恐怖袭击数量增多，这一类恐袭往往会选择卡车等交通工具作为工具。武器类型中基尼系数最低的类别是燃烧武器为 0.59，表明该变量聚集特征不太明显，即燃烧武器作为

恐袭工具在全球各个地区均被广泛使用。

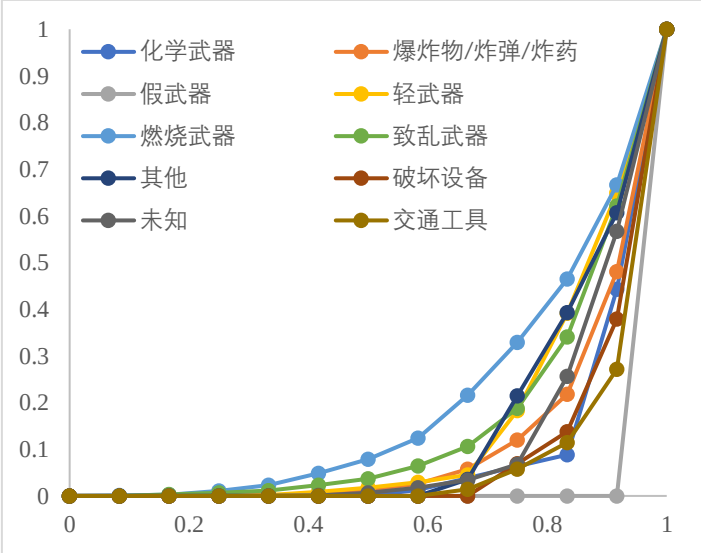

图 6.11 不同地区各武器类型的 Lorenz 曲线

表 6.4 不同地区各武器类型的基尼系数

武器类型	基尼系数
假武器	0.916667
交通工具	0.840476
破坏设备	0.818966
化学武器	0.809072
爆炸物/炸弹/炸药	0.76321
未知	0.757976
其他	0.708333
轻武器	0.695258
致乱武器	0.683042
燃烧武器	0.589283

2) 目标或受害者类型的空间聚集性

近三年在全球恐怖袭击事件中的目标或受害者类型可划分为 21 类，基尼系数均大于 0.6，聚集性普遍较强。其中基尼系数大于 0.8 的有 5 类，分别为流产有关、恐怖分子/非州立民兵组织、暴力政党、电信以及海事。以流产有关为目标的恐袭仅 7 起且全部发生于美国，且动机多为反对堕胎，这可能是因为美国社会对于堕胎是否合法长期存在较大争议。目标为恐怖分子或非州立民兵组织以及暴力政党和海事的恐怖袭击的热点地区均为中东和北非以及南亚，这些地区由于民族和宗教矛盾、社会发展不平衡以及西方势力介入等多种因素往往政局动荡、社会不稳定，所以各种恐怖组织及非法武装组织活动较为猖獗。以电信为袭击目标的恐袭热点地区为南亚和东南亚地区，笔者分析可能是因为南亚和东南亚地区人口比较密集，因此选择攻击这些地区的电信设施可能会造成比较大程度的混乱，更易达成恐怖分子的目的。而各分类中游客的基尼系数最小，

因为攻击游客的恐怖袭击发生往往比较随机，因此地区聚集性并不明显。

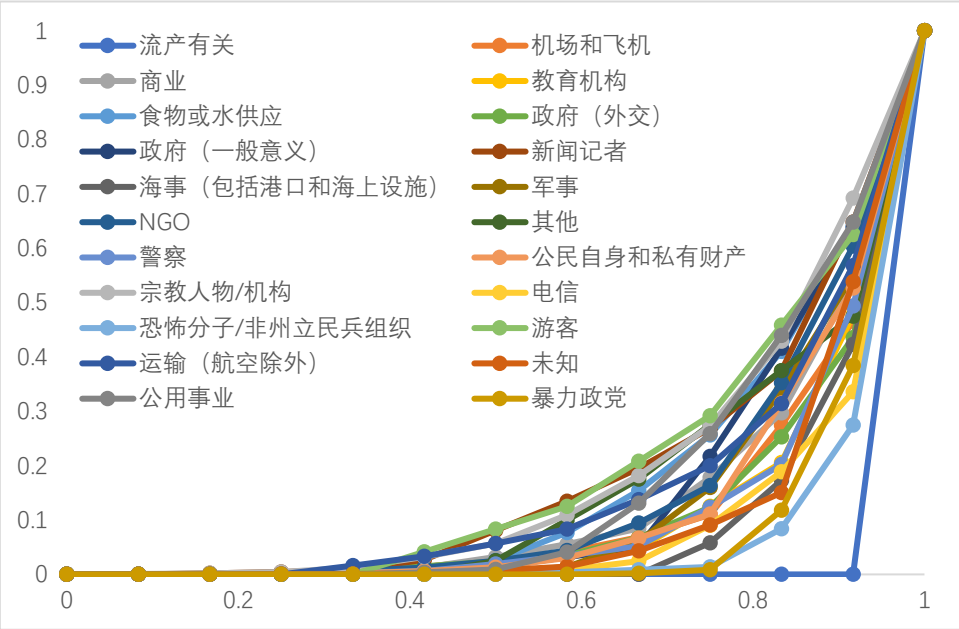

图 6.12 不同地区各目标或受害者类型的 Lorenz 曲线

表 6.5 不同地区各目标或受害者类型的基尼系数

目标或受害者类型	基尼系数
流产有关	0.916667
恐怖分子/非州立民兵组织	0.852627
暴力政党	0.831267
电信	0.808743
海事（包括港口和海上设施）	0.807692
未知	0.775119
教育机构	0.760671
警察	0.760448
政府（外交）	0.755882
机场和飞机	0.75
公民自身和私有财产	0.737987
军事	0.728641
商业	0.716102
NGO	0.699686
政府（一般意义）	0.686898
运输（航空除外）	0.681788
其他	0.679167
公用事业	0.661417
食物或水供应	0.655983
新闻记者	0.625934
宗教人物/机构	0.617736
游客	0.611111

3) 攻击类型的空间聚集性

近三年来恐怖袭击的攻击类型可划分为 8 类，基尼系数最大的类型为轰炸或爆炸，热点地区为中东和北非、南亚以及撒哈拉以南的非洲，这可能由于轰炸或爆炸恐袭事件发生较多，在总体中占比超过 50%，因此会与恐袭事件总体地区热点保持一致。攻击类型分类中基尼系数最小的是设施或基础设施攻击，表明这种攻击类型在全球范围内发生比较普遍。

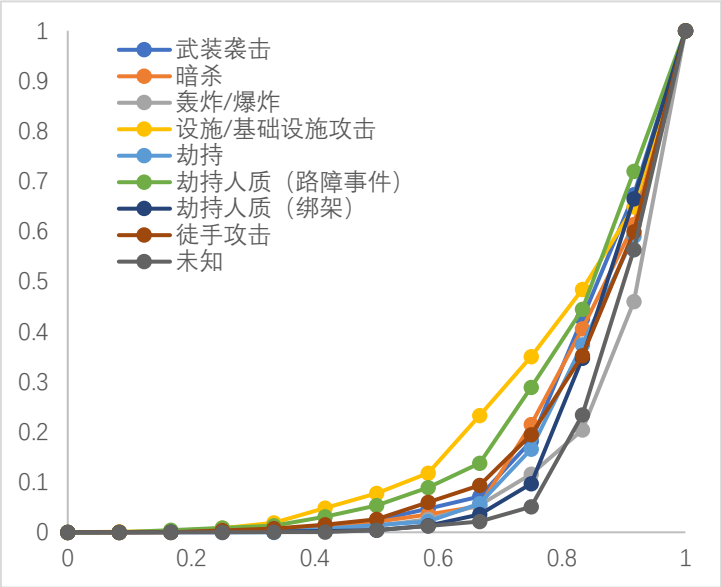

图 6.12 不同地区各攻击类型的 Lorenz 曲线

表 6.6 不同地区各攻击类型的基尼系数

攻击类型	基尼系数
轰炸/爆炸	0.770292344
未知	0.768764215
劫持人质（绑架）	0.722040328
劫持	0.711630695
徒手攻击	0.691323346
暗杀	0.690327613
武装袭击	0.676332782
劫持人质（路障事件）	0.618148148
设施/基础设施攻击	0.585037175

6.4 结论与建议

通过对近三年发生的全球恐怖袭击事件进行分析，本文主要给出以下几点结论与建议：

(1) 近年来全球反恐工作的开展卓有成效，恐怖袭击数量和成功率逐年减少，应该及时总结经验，巩固成果。

(2) 全球恐怖袭击事件在地域上分布十分不均衡，中东及北非、南亚和撒哈拉以南的非洲地区面临的反恐形势十分严峻。而从国家层面来看，恐怖袭击发生风险较高的国家有伊拉克、阿富汗、印度、巴基斯坦、菲律宾、尼日利亚、索马里、也门、埃及、叙利亚、利比亚、土耳其、泰国和乌克兰。

(3) 恐怖组织在全球的活动空间总体受到打压，但个别区域有蔓延风险。近年来东欧地区和东亚地区的恐怖主义活动遭到有效打击，但中美和加勒比海地区的恐怖主义开始逐渐抬头，北美地区则面临恐怖主义进一步扩散的风险。

(4) 无论是从武器、攻击目标还是攻击类型来看，全球恐怖袭击事件的特征均具有一定的空间聚集性。因此各国在制定反恐计划时既要考虑恐怖主义在全球的总体发展态势又要因地制宜、充分考虑本国与所在地区的实际情况。例如西欧地区应该重点关注常使用卡车等交通工具发动袭击的“独狼式”恐怖袭击，东南亚及南亚地区应加强本国通信设施的安全保障等。

七、问题四的建模与求解

7.1 恐怖袭击死亡人数的影响因素研究

7.1.1 问题提出

由第一问所得的结论可知，在恐怖袭击事件中，死亡人数成为每次事件危害性权重最高的指标，GTD 恐怖袭击事件已超 20 万起，死亡人数超 27 万人。

2017 年，恐怖袭击事件达到 10900 起，死亡人数和受伤人数分别为 26445 人和 24927 人，相比 2016 年分别降低 24%和 37%。全球恐怖袭击发展趋势见图。可见，恐怖袭击起数和伤亡人数在 2002 年至 2011 年间保持相对稳定，死亡人数和受伤人数在 2007 年和 2014 年达到了一次高峰，随后到 2015 年以后呈下降趋势；2002 年到 2011 年，伤亡人数处于稳定期，仅在年出现小的增长。由此可见，2011 年之后，恐怖袭击起数、死亡人数和受伤人数均呈现出了快速增长，并在 2014 年达到历史最高值。恐怖袭击表现出高致死率的趋势，而死亡人数越高，造成的社会影响和恐怖氛围越严重，因此，挖掘与高死亡水平相关的影响因素，有针对性的遏制恐怖袭击、降低人员伤亡显得迫切且重要。

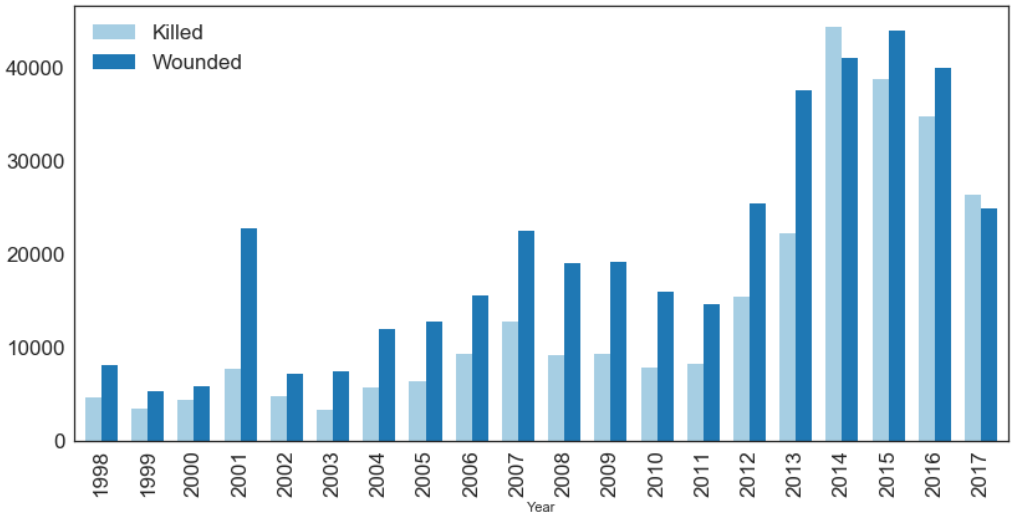

图 7.1　1998-2017 年死亡人数和受伤人数变化趋势

如果可以从数据中找到导致高死亡的关键因素，从而协助反恐活动制定相关决策，重点从某几个重要因素去尽量控制每次事件的死亡人数。因此，可以从以下几个维度去探究死亡人数的特征：死亡人数主要受哪几个因素影响？如何确定极端死亡人数的影响因素？何种武器或者袭击方式容易导致高死亡？

因此基于附件 1 中提供的数据，建立有监督的机器学习融合模型 lightgbm，结合多分类 logistic 串行融合，在保证模型效果的同时，还能够得到各个变量重要性，以及相互之间是如何影响的。

7.1.2 模型构建：lightgbm-logistic 串行模型求解

近 20 年来，每起恐怖袭击事件死亡人数不等，将近一半的事件没有死亡人

员，但是较严重的伤亡人数达到 1570 起。 为了使死亡的损害程度更有区分度，本章节将死亡人数按照以下几个等级分：

表 7.1 死亡人数区间划分

等级	死亡人数	频率
无死亡	0 人	52356
低死亡	1-3 人	40635
较低死亡	4-8 人	10076
高死亡	9-20 人	4753
大规模死亡	20 人以上	2083

（2）lightgbm

1）Lightgbm 模型是一种基于回归树的提升学习方法，归属于 boosting 融合方法。Lightgbm 在 xgboost 的基础上进行了较大的改进，相较于 xgboost 和 gbdt 又有明显优点，尤其在数据量非常大时，lighgbm 以其极低的内存消耗和远超 xgboost 的运算速度计算，大大提升了模型效率：

- Lightgbm 中的决策树子模型是采用按叶子分裂的方法分裂节点的，因此它的计算代价比较小，需要控制树的深度和每个叶子节点的最小数据量，从而可以防止模型过拟合。

2）Lightgbm 选择了基于 Histogram 的决策树算法，将特征值分为很多个小箱，进而在这些箱上寻找分裂，以此减小储存成本和计算成本。另外，类别特征的处理，也使得 Lightgbm 在特定数据下有比较好的提升。

3）Lightgbm 分为三类：特征并行、数据并行和投票并行。特征并行运用在特征较多的场景，数据并行应用在数据量较大的场景，投票并行应用在特征和投票都应用在比较多的场景。

因此，总结该算法的优点显著体现在如下五个方面：①更快的训练速度；②更低的内存消耗；③更好的模型精度；④支持并行学习；⑤可以快速处理海量数据。

此外，Lightgbm 中参数较多， 调整起来比较复杂， 所设置的参数会对预测结果产生显著影响， 对此我们在查阅相关资料后， 给出的调参方法如下:

STEP1. 设置初始值；

STEP2. 确定学习速率(eta)和相应的树的数目(nround)；

STEP3. 栅格搜索(grid search)确定每棵树的最大深度(max_depth)和最小节点权重(min_child_weight)；

STEP4. 调节 gamma 参数；

STEP5. 栅格搜索确定每棵树训练的样本子集(subsample)和特征子集(colsample_bytree)的大小；

STEP6. 栅格搜索确定每棵树的正则化参数 alpha 和 lambda；

STEP7. 重新设置更小的学习率， 确定此时的 nround。

（3）多分类逻辑回归模型

逻辑回归作为广义线性回归中的一种，其本身就具备可靠的统计理论基础，稳健性较高。由于其属于白箱模型，从而能提供各指标与违约情况的显示表达式，可以直观地反映各指标对违约的解释性，进而给出相应的经济学解释。

二分类模型的基本公式如下：

$$\text{logit}(\text{p}) = \text{X}\beta + \epsilon,\ \epsilon \sim \text{N}(0, \text{I}\sigma^2) \tag{7.1}$$

$$\hat{p} = \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})},\ \hat{p}\text{为违约概率} \tag{7.2}$$

上式称为 sigmod 函数，但是在过抽样的样本上训练模型时，预测概率会产生偏置$\ln(\frac{\rho_1\pi_0}{\rho_0\pi_1})$，此时模型 为：

$$\text{logit}(p^*) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + \text{X}\beta + \epsilon,\ \epsilon \sim \text{N}(0, \text{I}\sigma^2) \tag{7.3}$$

其 中，p^*是有偏样本的后验概率，并且有$\rho_0 < \pi_0$和$\rho_1 > \pi_1$，故调整后的后验概率为

$$\bar{p} = \frac{p^*\rho_0\pi_1}{(1-p^*)\rho_1\pi_0+p^*\rho_0\pi_1} \tag{7.4}$$

而在多分类逻辑回归模型中，可以将 sigmod 函数扩展成为 softmax 函数：

$$\hat{p}_i = \frac{\exp(X\hat{\beta}_i)}{\sum_i \exp(X\hat{\beta}_i)} \tag{7.5}$$

Softmax 函数得到的是一个[0,1]之间的值，且，这个 softmax 求出的概率就是真正的概率，换句话说，这个概率等于期望。

（3）lightgbm+logistic 串行模型

在串行结构的组合模型中，各分类器的学习是按照顺序进行的，前一个分类器的输出作为后一个的输入。如果能选择合适的分类器进行串行组合，那么组合分类器可能比单一分类器具有更加优秀的性质。

机器学习 lightgbm 模型分类精度高，但结果缺乏稳健性和解释性，传统的逻 辑回归模型尽管在分类精度上比机器学习模型稍显逊色，但其稳健性好，可解释 性高。所以，一种自然的解决思路就是将二者结合起来，形成优势互补，最终构 建一个分类精度、稳健性俱佳的组合模型。

根据这个思路，可以构建 lightgbm -logistic 串行组合模型：

图 7.2　lightgbm+logistic 串行模型框架

7.1.3 问题求解

由于死亡人数是事后指标，而在探究对其影响时，应该用事前指标，因此将后果一系列指标从数据中删除，并提取出有意义且缺失值较少的数据，借助python 中的 sklearn 模块对上述模型进行求解，得到结果如下：

表 7.2　死亡人数影响因素

变量	无死亡	低死亡	较低死亡	高死亡	大规模死亡
是否自杀式攻击	0.199593	0.933216	1.339053	1.322678	1.20546
国际意识形态	0.752449	0.770327	1.097432	1.225575	1.154217
ISIL	0.724753	0.93959	1.060259	1.163905	1.111494
是否声称	0.709855	0.804759	1.228812	1.163218	1.093355
撒哈拉以南的非洲	0.697811	0.987603	1.104166	1.143005	1.067415
Boko Haram	0.834148	0.956465	1.104933	1.055144	1.049309
是否持续性事件	1.127421	0.9056	0.947725	0.983649	1.035605
国际后勤	1.082217	0.860513	1.00548	1.016188	1.035602
是否绑架	1.110409	0.986229	0.926403	0.958661	1.018299
政治	0.9872	1.030419	0.976462	0.989351	1.016568
平民	0.693659	1.077831	1.177456	1.035648	1.015406
Al-Qaida in Iraq	0.936797	1.006003	1.06038	0.986199	1.010622
绑架人数	0.9985	0.995075	0.998939	0.999906	1.00758
凶手个数	1.002555	0.990186	0.99797	1.00348	1.005808
释放人质数	0.999422	0.995622	0.998702	1.00059	1.005664
Al-Shabaab	1	1	1	1	1
是否个体作案	1.024411	0.998473	0.982299	0.995169	0.999649
其它	1.002883	0.999964	1.000688	0.997709	0.998756
其它	1.004168	1.005056	0.995202	0.997239	0.998335
亚洲其它地区	1.035524	0.998358	0.981684	0.987052	0.997382
劫持人质	1.244278	0.940064	0.890155	0.930215	0.995288
相关的事件数	1.028392	0.990059	0.996887	0.992133	0.992529
是否邻近城市	1.013791	0.971614	1.018614	1.004205	0.991776
是否事件组	1.215583	0.713125	1.027234	1.054529	0.989528

由上表可得到以下结论：

大规模死亡事件影响因素：自杀式袭击、国际意识形态、ISIL 恐怖组织、Boko Haram 组织、撒哈拉以南的非洲、是否持续性事件。

由此表明，恐怖袭击组织一旦用自杀式袭击方式发起恐怖事件，将造成不可想象的后果，911 事件客机直接撞向大楼等自杀式方式直接造成 3000 多人的损伤，5 座大楼的毁损；因此，反恐政策可以从加强保护客机、轮船等可能出现自杀式袭击的场景的保护以及相关武器的限制。

组织者与被害者是否同一国籍也可能对大规模死亡造成较大的影响，因此国防相关部门应该加强维护，同时严格对海外人员进行审查，尽量减少组织海外入境的概率。

此外，ISIL 恐怖组织和 Boko Haram 组织发起的恐怖袭击事件也很可能会造成较大的损伤，因此反恐工作者还应该重点观察这两个组织的动向，尽量能够提

前识破其蓄谋的恐怖袭击事件。

相对来说，对于恐怖袭击造成的死亡人数为 0 的事件中，威胁、恐吓群众、政治、宗教目标等因素系数较大，说明其对死亡人数的影响影响相对较小，往往这种事件并非蓄谋造成人身伤害。而在低死亡人数的恐怖袭击事件中，Taliban 组织、是否声称负责、武装袭击影响较大，在日常反恐工作中，应该尽可能将低死亡人数控制为无死亡，可以通过掌握相关恐怖组织（Taliban） 信息来有效防范恐怖袭击事件的发生。

7.2 恐怖袭击事件数增长率的 ARMA 干预模型

7.2.1 问题提出

恐怖主义袭击事件发起本身可能具有某种规律，亦或许会有某种新的特征和新的趋势。有数据表明，国际恐怖主义活动的演进可能具有某种周期波动的特征，即当某一时期增长较快时，下一时期的国际恐怖活动的活跃程度就会下降，反之，国际恐怖主义活动就会再度兴起。

这里以九一一事件为例，九一一事件是历年来国际恐怖袭击事件中危害性较高的事件之一，自此之后，国际社会和世界各国迅速做出反应，纷纷加大反恐力度，给国际恐怖组织以沉重打击。而国际反恐趋势的加大力度必然同时也会使恐怖袭击事件有新动向。通过探究九一一事件后国际反恐态势的新特征和新趋势，有助于有针对性地完善现有反恐政策。

ARMA 干预模型能够有效控制事物自身所带来的内生影响，非常适合分析特定外部冲击所引起的新变化。因此，我们利用附件 1 中的数据，对一定时期内的恐怖袭击事件数增长率和伤亡人数增长率建立 ARMA 干预模型，探索其自身增长规律，验证现有成果所提出国际恐怖主义发展的部分新特征，同时能够获得其他方面的发展趋势。

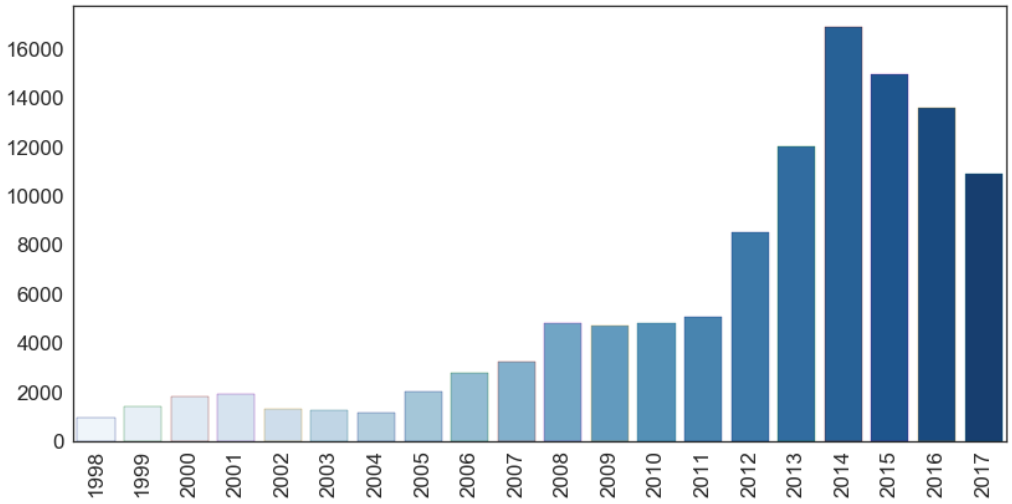

图 7.3 1998-2017 年恐怖袭击事件起数

1998 年-2014 年，恐怖袭击事件数逐年增高，死伤人数也逐步提高，到 2014 年，恐怖袭击事件数 16903 件，死亡人数 44490 人，受伤人数 41128，均历史高点。2014 年以后，恐怖袭击逐步得到遏制，事件发起次数与伤亡人数都在逐年降

低。此外，从图中可以明显看到 2001 年恐怖袭击相比前后几年都较多，伤亡人数也较前一年较高的增长，主要受到 911 事件的影响，全球反恐态势有了新的特征。

7.2.2 模型的构建：ARMA 干预模型

（1）ARMA 干预模型

ARMA 干预模型最早由由乔治·博克斯和乔治·泰奥[12]在 1965 年提出，该模型能够有效控制事物自身所带来的内生影响，因此适合分析特定外部冲击所引起的新变化，对事件影响和政策效果进行评估。理查德·麦克利里、理查德·海进一步说明模型由两部分组成: 一部分是由 ARMA 组成的“噪声部分”; 另一部分则是由虚拟变量构成的“转换方程”。

使用传统的方法估计 ARMA 干预模型对数据数量的要求较高，一般需要 50~100 个样本。当面临样本较少的情况时[13]，可以借鉴约翰·戈特曼(John M． Gottman) 提出的 ITSE 方法对其进行改进,样本数约 20 个即可。附件 1 中提供了 1998~2017 年的恐怖袭击事件，样本量刚好 20。

基于 ITSE 方法和 ARMA 干预模型的噪声部分（N_t）和转换方程（F_t）表达式：

$$N_t = \sum_{i=1}^{p} a_i y_{t-i} - \sum_{j=1}^{q} b_j u_{t-j} + u_t \tag{7.6}$$

$$F_t = cv_t + dv_t T_t \tag{7.7}$$

从而，整个干预模型为:

$$y_t = N_t + F_t = \sum_{i=1}^{p} a_i y_{t-i} - \sum_{j=1}^{q} b_j u_{t-j} + cv_t + dv_t T_t u_t \tag{7.8}$$

其中 y_t是自回归序列，表示第 t 年由于国际恐怖主义袭击导致死亡人数的增长率。u_t是相互独立的白噪声序列，且服从均值为 0、方差为σ_u^2的正态分布。

v_t是虚拟变量，衡量九一一事件是否已经发生。

$$v_t = \begin{cases} 1, & t > 2002 \\ 0, & t < 2001 \end{cases} \tag{7.9}$$

T_t 是时间趋势变量，a_i、b_j、c 和 d 都是待估计的参数。

---

[12] Adam Z.Rose and S.Brock Blomberg，“Total Economic Consequences of Terrorist Attacks: Insights from9 /11，”Peace Economics，Peace Science and Public Policy，No.1，2010:1154 － 1189．

[13] Rousseau Cecile，Hassan Ghayda，Moreau Nicolas and Brett D． Thombs，“Perceived Discrimination and Its Association with Psychological Distress among Newly Arrived Immigrants before and after September 11，2001，” American Journal of Public Health，No.5，2011: p.7．

八、模型评价与改进

问题一是对恐怖袭击事件划分伤害等级。考虑到恐怖袭击危害程度受诸多因素影响，因此本文通过提取一系列与之相关的特征，采用因子分析对提取的特征进行合理地降维处理，让得到的公共因子对危害程度具有更好解释性的同时，也大大简化了工作量。接着，在对因子赋予权重这一步，为了得到更贴近真实值的权重，本文采用层次分析和序关系分析这两类主观赋权法，并结合两类客观分析法，根据矩法估计理论，形成主客观组合赋权的优化模型，输出各因子的综合权重，由此计算出的综合得分，采用灰色综合评价法的关联分析，并结合使用 k-means 聚类方法输出最终的危害等级。根据本文创建的恐怖袭击危害评分与等级划分体系，可以输出近二十年来的十大恐怖袭击事件，其中位于最前两位的正是发生于美国纽约的 911 事件，由此可见，本文创建的主客观评分体系具有较高的现实意义。

问题二是根据事件特征寻找嫌疑人的问题，首先采用 OPTICS 算法对事件进行聚类，OPTICS 具有聚类簇的形状没有偏倚、不需提前设定聚类个数且对参数设置不太敏感等优点，聚类结果可信度高。之后对每个恐怖组织或个人的危害程度进行量化，从恐怖袭击发起数量和危害程度两个层面考虑设计了评分模型。最后利用 XGBoost 算法对事件嫌疑人进行判别，XGBoost 模型整体表现良好。后续还可以考虑采用 stacking 等方法进一步提升 XGBoost 的模型效果。

问题三是对未来全球反恐态势研判的问题，首先采用数据可视化等技术分别从时间、空间以及时间加空间角度对今年以来全球恐怖袭击事件发生特点、蔓延特性、级别分布等方面进行了较为全面的分析。之后将 Lorenz 曲线、基尼系数与帕累托分析法结合研究了不同变量的空间聚集性与热点，并发现国家与地区等区域因素的空间聚集性十分显著，基于此进一步从不同角度出发挖掘了恐袭事件在不同区域的特点。总体而言对目标问题的考察较为全面与深入、并且主次分明。但本文并未分析除时间空间以外的各变量之间的交互影响，之后可以考虑从这一角度出发进行探究。

问题四是对附件数据的进一步拓展。本文重点观察了恐怖袭击事件中的每年事件数与死伤人数，并提出了建立 lightgbm-multilogistic 串行模型以挖掘影响恐怖袭击中影响死伤人数的主要因素，以及基于 ITSE 方法对每年恐怖袭击事件数建立 ARMA 干预模型以探究 911 事件后全球恐怖袭击态势变化及反恐效果。总体而言，找到了合适且有价值的问题去分析研究，但是附件中仍有大部分数据的价值没有被开发，后期或许可以从更细节的角度去发掘（例如事件发生具体地点、动机等等）。

参考文献

[1]郭金玉, 张忠彬, 孙庆云. 层次分析法的研究与应用[J]. 中国安全科学学报, 2008, 18(5):148.
[2]李连结，姚建刚, 龙立波等．组合赋权法在电能质量模糊综合评价中的应用[J]．电力系统自动化，2007,31(4)：56-60．
[3]聂宏展,方吕盼,乔怡等．基于熵权法的输电网规划方案模糊综合评价[J]．电网技术，2009,33(11):60-64.
[4]王应明,张军奎．基于标准差和平均差的权系数确定方法及其应用[J]．数理统计与管理，2003,22(7):22-26.
[5]杨虎,刘琼荪,钟波．数理统计[M], 北京:高等教育出版社，2004
[6]江文奇. 多属性决策的组合赋权优化方法[J]．运筹与管理，2006,15(6): 40-43.
[7]刘思峰, 等. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2010
[8]莫豪文. 数据挖掘方法在反恐预警中的应用[D]. 北京工业大学, 2017.
[9]邓灵评. 基于数据挖掘的犯罪行为分析及系统实现[D]. 西南交通大学, 2014.
[10]李国辉. 全球恐怖袭击时空演变及风险分析研究[D]. 中国科学技术大学, 2014.
[11] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996:226-231
[12] Ankerst M, Breunig M M, Kriegel H P. OPTICS:ordering points to identify the clustering structure[J]. Acm Sigmod Record, 1999, 28(2):49-60.
[13] Chen T, Guestrin C. XGBoost:A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.
[14] Adam Z.Rose and S.Brock Blomberg，“Total Economic Consequences of Terrorist Attacks: Insights from9 /11，”Peace Economics，Peace Science and Public Policy，No.1，2010:1154 －1189．
[15] Rousseau Cecile，Hassan Ghayda，Moreau Nicolas and Brett D． Thombs，“Perceived Discrimination and Its Association with Psychological Distress among Newly Arrived Immigrants before and after September 11，2001，” American Journal of Public Health，No.5，2011: p.7．

附录

附录 A：综合得分前/后十样本及其等级

事件编号	综合得分	等级
200109110005	28.35165	1
200109110004	28.34741	1
200107240001	18.66749	1
199808070002	11.43139	2
201406150063	8.93618	2
201602180049	5.35731	2
201412070129	5.317995	2
201710140002	4.877824	2
201406100042	4.314157	2
…	…	…
201711070042	-0.62589	5
200904220015	-0.63581	5
201505140080	-0.6445	5
201608120051	-0.64461	5
201505140079	-0.65695	5
200001240001	-0.73243	5
201609040013	-0.83691	5
200910270022	-1.02495	5
201707260012	-1.07043	5
201408230034	-1.81836	5

附录 B：code

```
##主观和客观方法的结合,求出最终的权重
from scipy.optimize import minimize
def fun(args): ##输入的参数代表常数
    alpha,beta,W1,W2,W3=args
    v=lambda w: beta*np.sum((np.array(w)-np.array(W1))*(np.array(w)-
np.array(W1))) \
                +beta*np.sum((np.array(w)-np.array(W2))*(np.array(w)-
np.array(W2))) \
                +alpha*np.sum((np.array(w)-np.array(W3))*(np.array(w)-
np.array(W3)))
    return v
def con(args):
    ##约束条件分为 eq 和 ineq
    ##eq 表示函数结果等于 0，ineq 表示 ineq 表示函数结果大于等于 0
```

```
a,b,c=args ##a:1 向量      b:-1 向量  c:1
cons=({'type':'ineq','fun':lambda w:np.sum(np.array(w)*a)-1},
        {'type':'ineq','fun':lambda w:np.sum(np.array(w)*b)+1},
        {'type':'ineq','fun':lambda w:w[0]},
        {'type':'ineq','fun':lambda w:c-w[0]},
        {'type':'ineq','fun':lambda w:w[1]},
        {'type':'ineq','fun':lambda w:c-w[1]},
        {'type':'ineq','fun':lambda w:w[2]},
        {'type':'ineq','fun':lambda w:c-w[2]},
        {'type':'ineq','fun':lambda w:w[3]},
        {'type':'ineq','fun':lambda w:c-w[3]},
        {'type':'ineq','fun':lambda w:w[4]},
        {'type':'ineq','fun':lambda w:c-w[4]},
        {'type':'ineq','fun':lambda w:w[5]},
        {'type':'ineq','fun':lambda w:c-w[5]},
        {'type':'ineq','fun':lambda w:w[6]},
        {'type':'ineq','fun':lambda w:c-w[6]},
        {'type':'ineq','fun':lambda w:w[7]},
        {'type':'ineq','fun':lambda w:c-w[7]},
        {'type':'ineq','fun':lambda w:w[8]},
        {'type':'ineq','fun':lambda w:c-w[8]},
        {'type':'ineq','fun':lambda w:w[9]},
        {'type':'ineq','fun':lambda w:c-w[9]},
        {'type':'ineq','fun':lambda w:w[10]},
        {'type':'ineq','fun':lambda w:c-w[10]},
        {'type':'ineq','fun':lambda w:w[11]},
        {'type':'ineq','fun':lambda w:c-w[12]},
        {'type':'ineq','fun':lambda w:w[13]},
        {'type':'ineq','fun':lambda w:c-w[13]},
        {'type':'ineq','fun':lambda w:w[14]},
        {'type':'ineq','fun':lambda w:c-w[14]},
        {'type':'ineq','fun':lambda w:w[15]},
        {'type':'ineq','fun':lambda w:c-w[15]},
        {'type':'ineq','fun':lambda w:w[16]},
        {'type':'ineq','fun':lambda w:c-w[16]},
        {'type':'ineq','fun':lambda w:w[17]},
        {'type':'ineq','fun':lambda w:c-w[17]},
        {'type':'ineq','fun':lambda w:w[18]},
        {'type':'ineq','fun':lambda w:c-w[18]},
        {'type':'ineq','fun':lambda w:w[19]},
        {'type':'ineq','fun':lambda w:c-w[19]},
        {'type':'ineq','fun':lambda w:w[20]},
        {'type':'ineq','fun':lambda w:c-w[20]},
        {'type':'ineq','fun':lambda w:w[21]},
```

```
                {'type':'ineq','fun':lambda w:c-w[22]},
                {'type':'ineq','fun':lambda w:w[23]},
                {'type':'ineq','fun':lambda w:c-w[23]},
                {'type':'ineq','fun':lambda w:w[24]},
                {'type':'ineq','fun':lambda w:c-w[24]},
                {'type':'ineq','fun':lambda w:w[25]},
                {'type':'ineq','fun':lambda w:c-w[25]})

    return cons

args=(0.5815,0.4185,W1,W2,W3)
#设置参数范围/约束条件
args1=(np.ones(num_factor),-np.ones(num_factor),1)      #x1min,  x1max,  x2min,
x2max
cons=con(args1)
#设置初始猜测值
w0=[1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
model=minimize(fun(args),w0,method='SLSQP',constraints=cons)
##结果
W_total=model.x

2.
##xgboost 模型输出给定 10 个事件的嫌疑人
import xgboost as xgb
from sklearn.cross_validation import train_test_split

# xgboost 参数
params = {'booster': 'gbtree',
                'objective': 'multi:softprob',
                'gamma': 0.025,      ##'objective': 'multi:softmax'      多分类，直接输出
概率最大的类别
                'min_child_weight': 6,
                'max_depth': 7,
                'lambda': 1,
                'subsample': 0.7,
                'colsample_bytree': 0.6,
                'eta': 0.1,
                'seed': 0,
                'max_delta_step': 0.5,
                'silent': 0,
                'scale_pos_weight': 4,
                'num_class': 5
                }
new_data2['suspect_id']=new_data2['suspect_id'].apply(lambda x:int(x))
```

```
train=xgb.DMatrix(new_data2[feature_list],new_data2['suspect_id'])
test=xgb.DMatrix(test_data2[feature_list])

model=xgb.train(params,train,num_boost_round=500)
test_id=model.predict(test).reshape(test_data2.shape[0],5)
id_prob=pd.DataFrame((-
1*test_id).argsort(axis=1)+1,columns=['id1','id2','id3','id4','id5'])
result=pd.concat([raw_test_data2[['eventid']],id_prob],axis=1)
```